

# Massive Data Sets: Challenges for Statistics

*Karen Kafadar*

*Indiana University*

*kkafadar@indiana.edu*

<http://math.cudenver.edu/~kk>

## **Acknowledgements:**

*E.J. Wegman, GMU; D. Marchette, NSWC; G. Davis, UVA;*

*R.G. Jacobsen, UC-Berkeley*

*NSF DMS-0527090; ARO W911NF0510490*

## OUTLINE

1. Motivation: Massive data from Genomics, Internet, High-energy physics; Fraud detection literature
2. Guiding principles; Take-home messages
3. Genomics: Microarray data; Multiple testing
4. Internet traffic data (Sample data from GMU)  
IP addresses, ports, message sizes
5. High-energy physics experimental data  
Classification, Tiny signals in massive background
6. Morals: Computing, Analysis, Inference
7. Further research  
New framework for inference from massive data sets?

## 1. Motivation: Massive Data

- Genomics:

Measure mRNA expression in response to stimulus from  
50K-70K genes/person

Goal: Identify function of genes

Size: 0.5–1.0 Million/experiment ( $50K \times 20$  people)

Relatively easy: Specified hypotheses

- Internet:

Cybersecurity, computer viruses, network attacks:  
Scientific computing, financial transactions, business  
operations, security procedures, ...

Goal: Detect attacks *before* they force shutdowns

Size: Thousands of packet transmits per minute

- **High-energy Physics (HEP):**

- Colliding beams of electrons (SLAC) or protons (CERN) accelerated at very high energies (MeV/GeV/TeV)
- Collisions yield short-lived particles that decay into more short-lived particles in any one of 100,000 ways (“events”)
- Most events well-characterized (particles, speeds, lifetimes)
- Others less well understood (e.g. those with B-mesons)

**Goal:** Find target events of interest amidst millions of “uninteresting” events

**Size:** Millions per minute

## Common Theme: Tiny signal in vast sea of noise

Detect abnormal behavior: disease surveillance; nuclear product mfg; phone/charge card usage; financial transactions; ...

Easier when:

- data streams stratify into smaller data sets
- smaller sets are roughly independent of each other
- sets can be modeled simply and parametrically
- nature of potential abnormality is well-characterized
- residual distribution is well specified  $\Rightarrow$  assess *probability* of abnormality
- SPC-type tools are applicable

## Specification of hypotheses

**Genomics:** Many hypotheses – but straightforward

**Internet:** Many vaguely-specified hypotheses (outliers, excess packets/transmissions, signals of potential attacks)

**HEP:** Innumerable partially-specified complex hypotheses

Comparisons of thousands of likelihoods is impractical

Likelihoods based on convenient model assumptions (Gaussian, independence, ...)

Most frequent events are well-understood

“Interesting” events occur only rarely (0.1%)

**Goal:** Reduce “background” (uninteresting or well-characterized) events and remove them (EDA)

### Features of Internet traffic data:

- Relentless (“streaming”)
- Not independent of other systems: thousands of messages from thousands of ports/addresses each *minute*
- Diverse (text, numeric, image)
- Dispersed (geographically)
- Data often not from some convenient mathematical pdf

What data should be collected?

How can anomalies be detected?

### Features of HEP data:

- Relentless (“streaming”)
- “Events” are *assumed* to be relatively independent of each other (e.g., occurrence of event type A tells us nothing about the probability of event B occurrence)
- High-precision measurements — when we can see them
- Mis-identified tracks associated with events
- Data are *not* from a convenient mathematical pdf



An example of an “inconvenient” mathematical pdf  
(from high-energy physics):

$$\begin{aligned}
 f(q^2 = (p_\ell + p_\mu)^2, \cos \theta_\ell, \cos \theta_\nu, \chi) = & \\
 \frac{3G_F^2 |V_c b|^2}{8(4\pi)^4} \frac{\rho_{D^*}}{(q^2 - m_l^2)^2} q^2 B \times [ & ((1 + \cos^2 \theta_\ell) + \frac{m_\ell^2}{q^2} \sin^2 \theta_\ell) \sin^2 \theta_V (|H_+|^2 + \\
 |H_-|^2) + 4(\sin^2 \theta_\ell + \frac{m_\ell^2}{q^2} \cos^2 \theta_\ell) \cos^2 \theta_V |H_0|^2 - 2 \cos \theta_\ell \sin^2 \theta_V (|H_+|^2 - & \\
 |H_-|^2) - 2(1 - \frac{m_\ell^2}{q^2}) \sin^2 \theta_\ell \cos 2\chi \sin^2 \theta_V \operatorname{Re}(H_+ H_-^*) + (1 - & \\
 \frac{m_\ell^2}{q^2}) \sin 2\theta_\ell \cos \chi \sin 2\theta_V \operatorname{Re}(H_+ + H_-) H_0^* - & \\
 \sin \theta_\ell \cos \chi \sin 2\theta_V \operatorname{Re}(H_+ - H_-) H_0^* + 4 \frac{m_\ell^2}{q^2} \cos^2 \theta_V |H_t|^2 &
 \end{aligned}$$

(Can be simplified by assuming  $H_-$ ,  $H_0$ ,  $H_+$  are real)

Graphics and visualization are critical

## Fraud Detection Literature

- *General overview*: R.J. Bolton, D.J. Hand (2002), “Statistical fraud detection: A review” (with discussion), *Stat. Sci.*
- *Experimental design*: M. Schonlau, W. DuMouchel, W-H Ju, A.F. Karr, M. Theus, Y. Vardi (2002), “Computer intrusion: Detecting masquerades,” *Stat. Sci.*: designed experiments to evaluate algorithms for detecting masquerade user (stratify by user, identify characteristic features of user’s “signature”)
- *Modeling*: W.S. Cleveland, D.X. Sun (2000), “Internet traffic data,” *JASA*: Models for times between web accesses and challenges of long-range dependence and stationarity

- *Telephone calling fraud* (stratify by user): K.C. Cox, S.G. Eick, G.J. Wills (1997), “Visual data mining: Recognizing telephone calling fraud;” C. Cortes, D. Pregibon (2001), “Signature-based methods for data streams,” *Data Mining and Knowledge Discovery Data Mining and Knowledge Discovery*:
- *Nuclear product manufacturing*: Spiegelman+Rosenblatt 1984
- *Disease surveillance*: Siegrist et al. 2004, Stroup et al. 1989, Waller and Gotway 2004
- *Bioterrorism*: Hutwagner et al. 2003, many others
- *Visualizing network data*: S. Krasser et al. 2005: “Real-time and forensic network data analysis using animated and coordinated visualization” *IEEE Workshop on Information Assurance, USMA*: PC/time plots

New data types/structures have lead to advances in statistics  
(EJW, PJH)

- Data from agricultural expts  $\Rightarrow$  Design of experiments
- “Large” data sets  $\Rightarrow$  Statistical graphics
- No specified probability distribution  $\Rightarrow$  Nonparametrics
- ‘Almost’ Gaussian distributions  $\Rightarrow$  Robust methods
- Many-featured data  $\Rightarrow$  Multivariate statistics/displays
- Clinical trials  $\Rightarrow$  Sequential analysis
- Testing many hypotheses  $\Rightarrow$  Multiple comparisons
- Many other examples ...

## “Take-Home Messages”

- Prevalence of *streaming* data will increase
- Basically unusable in raw form; require much pre-processing
- Detecting “exotic” requires characterizing “typical”
- New challenges for interdisciplinary teams:
  - data value: what data to collect/discard
  - data warehouse: acquisition, storage, distribution
  - tools, algorithms for pre-processing
  - data analysis: robust, efficient, “sufficient”
  - informative visual graphical displays
  - automated interpretation of visual inferences (P.K. Banerjee,  
‘Automated band detection in remotely sensed imagery’)

## Guiding Principles

### Lessons from EDA

*“... ‘exploratory data analysis’ is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as for those we believe might be there. Except for emphasis on graphs, its tools are secondary to its purposes.... the tool-kit of exploratory data analysis is, and must remain, open-ended.”*

*“Data analysts regard their models as a basis from which to measure deviation, as a convenient bench mark in the wilderness, expecting little truth and relying on less.”*

– JWT, “Comment” (Parzen), *JASA* 1979, pp.121-122

*“Statistics is ‘reactive’ — very responsive to new problems that arise in chemistry, biology, physics, ...” — P. Hall*

*“Advances in powerful computing equipment has had a dramatic impact on statistical methods and theory. It has changed forever the way data are analyzed” – P. Hall*

*“Far better an approximate answer to the **right** question, which is often vague, than an **exact** answer to the wrong question, which can always be made precise” – JWT 1962, p.13*

*“Better an **imprecise** measure of something **important** than a **precise** measure of something **unimportant**” – D. Byar*

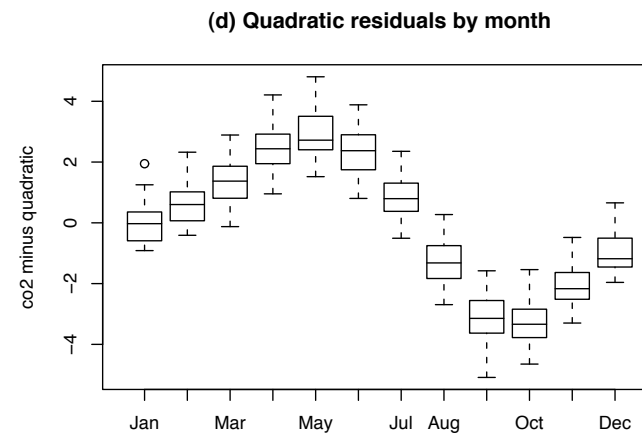
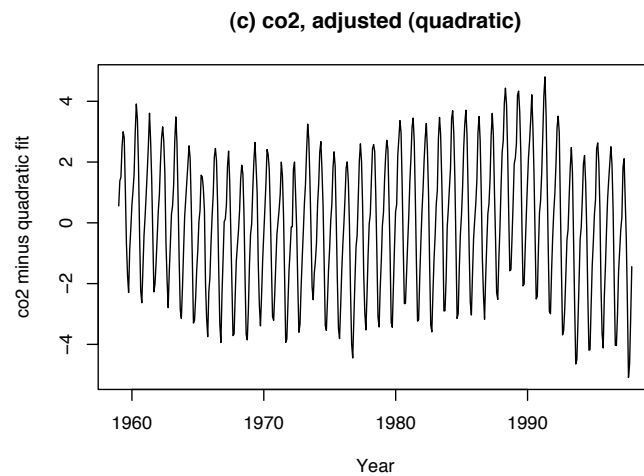
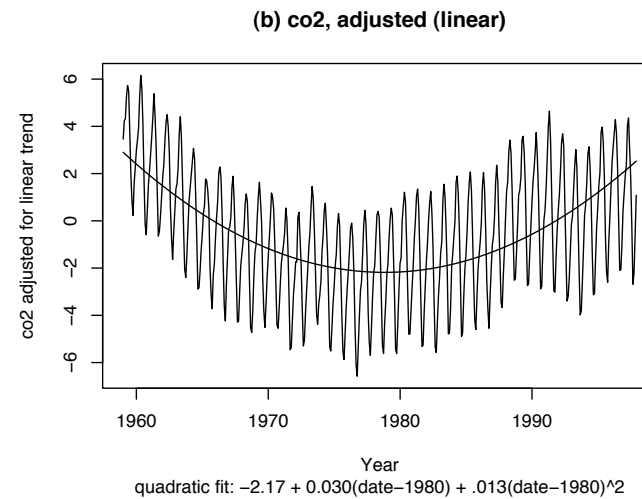
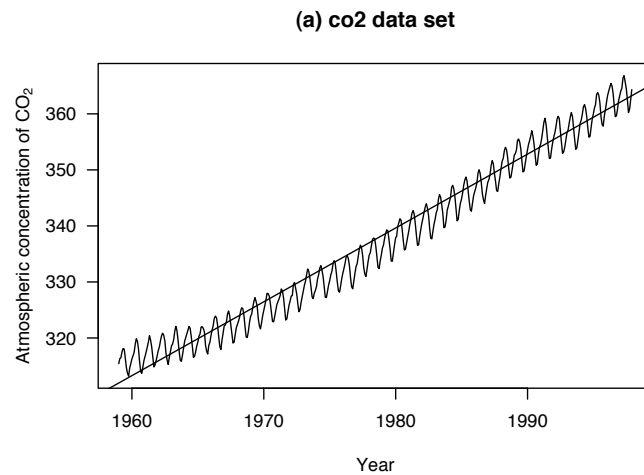
*“The greatest value of a picture is when it forces us to notice what we never expected to see” – JWT, *EDA*, p.vi*

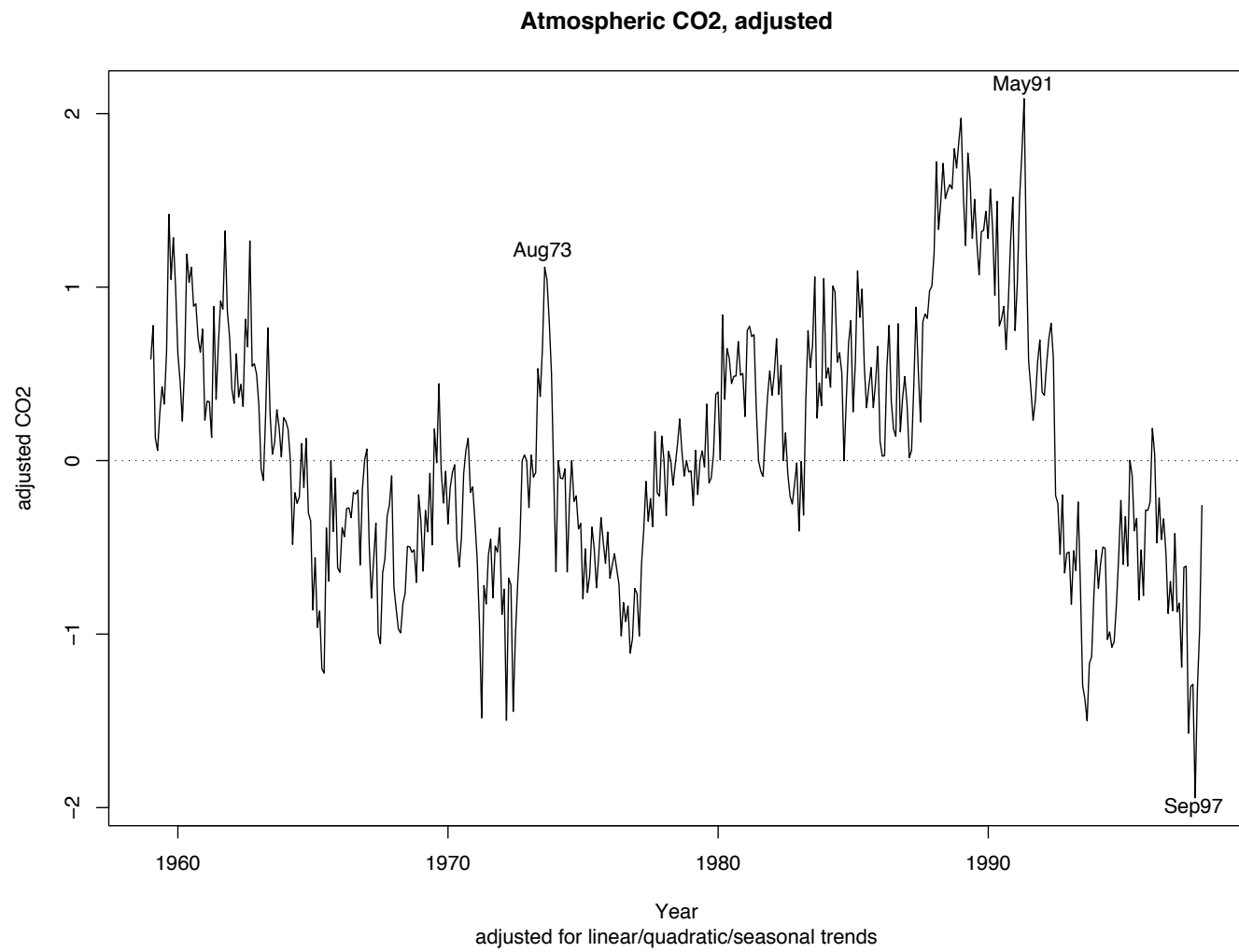
“R<sub>x</sub>”:

- Start with simplest of models
- Remove the obvious, magnify the residual
- Iterate robustly, note non-conforming pieces
- Control  $E(\textit{missed})$ , not  $P\{\textit{missed}\}$  — by piece
- Make good use of graphical displays
- “Cognostics”, “scagnostics” – diagnostics from cognitive/scatterplot displays (Tukey + Tukey 1985)



*“Graphs will certainly be increasingly ‘drawn’ by the computer without being touched by hands. More and more, too, as better procedures of diagnosis and indication are automated, graphs, and other forms of expository output, will, in many instances, cease to be the vehicle through which a man diagnoses or seeks indications, becoming, instead, the vehicle through which the man supervises, and approves or disapproves, the diagnoses and indications already found by the machine.” – JWT 1962, p.60*





### 3. Microarray data

- DNA (genetic code: A,C,G,T) in cell **nucleus**
- triplets of nucleotides code for **amino acids**  
(ex: AAA or AAG codes for Lysine)
- Genes = organized strings of nucleotides (in triplets)
- Genes code for proteins (e.g., insulin)
- Proteins made in cell **cytoplasm**
- Cell needs genes into cytoplasm (only coding part of DNA)
- DNA copy (cDNA: AT/CG); Excise introns; T  $\rightarrow$  U (mRNA)
- **Gene expression = measure of mRNA concentration** —  
which *may* correlate with protein production

Ex (Kim Kafadar):  $\text{Ca}^{2+}$  signaling in yeast cells

$\text{Ca}^{2+}$  needed for cells to survive stresses (e.g., high salt, alkaline pH, cell wall damage); promotes signaling through calcineurin (protein phosphatase)

*What does calcineurin do?*

dephosphorylates & activates Crz1p/Tcn1p/Hal8p transcription factor; Crz1p accumulates in cell nucleus, activates gene transcription whose products promote adaptation to stress

*What inhibits Crz1p?*

Start with  $\text{Ca}^{2+}$

## cDNA Experiment Preparation

- Grow 2 batches of cells: w/ $\text{Ca}^{2+}$ , w/o  $\text{Ca}^{2+}$
- Wait 30 ( $\pm$ ) min; harvest cells, collect mRNA
- Reverse-transcribe mRNA back to cDNA ( $\text{U} \rightarrow \text{T}$ )
- Denature (“un-zip” — single strand)
- Label “no-Ca” cells w/green flourophore (532nm)
- Label “Ca” cells w/red flourophore (635nm)
- Hybridize both to cDNA slide

Process variability at each step

cDNA slide:

- Single-strand copy of each gene (“probe”) printed in defined locations (“spots”)
- Ca/no-Ca fluorophore-labeled cDNA mixture placed on slide
- cDNAs in mixture find matching partners
- match  $\Rightarrow$  binding energy  $\Rightarrow$  radiates (532 if green; 635 if red)
- Laser scanner records fluorescence by pixel
- “Red” if more red-tagged mRNA (Ca batch)
- “Green” if more green-tagged mRNA (no-Ca batch)
- “Yellow” if not much difference

*Which spots (genes) have more red than green? (Multiplicity)*

- Bonferroni too conservative ( $\alpha/N$  when  $N = 50,000$  is  $10^{-6} \Rightarrow z_{crit} = 4.89$ )
- MC procedures control  $P\{\geq 1 \text{ False Positive}\} \leq \alpha$
- Benjamini & Hochberg (JRSSB 1995): FDR  
Control *expected proportion of false positives*  $\leq \alpha$   
( $W = \#$  significant by chance alone  $\sim Bi(N, \alpha)$ ;  
 $E(W)$  (FDR) easier than quantile  $W_{1-\alpha}$ )

Simplified: slides “spot” genes in “blocks” of  $\sim 500 - 1000$

(red, green) highly-correlated data pairs  $\sim$  3-parameter lognormal;  
transform to bivariate Gaussian  $N_2(0, I)$ ; look for genes outside  
circle

**Process issues; Big  $p$ , small  $n$ ; correlation among  $p$  (genes)**



Actual data values: For each channel (532nm, 635nm):

- # of foreground (spot) pixel intensities
- Diameter of foreground spot
- # of background pixel intensities
- median, mean, SD of foreground pixel intensities
- median, mean, SD of background pixel intensities

Which pixels are used to compute the local background? First, a circular region is drawn that is centered on the feature-indicator. This region has a diameter that is three times the diameter of the current feature-indicator. All of the pixels within this area are used to compute the background unless one of the following is true:

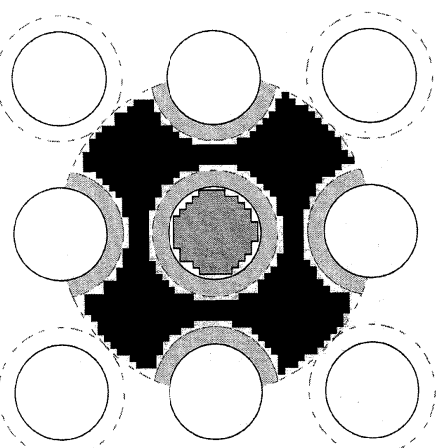
- the pixel resides in a neighboring feature-indicator;
- the pixel is not wholly outside a two pixel wide ring around a feature-indicator;
- the pixel is within the feature-indicator of interest.

In Figure 1, the black region represents the pixels used for computing the background, the dark gray region represents the pixels used for the feature intensities, and the light gray region represents excluded pixels.

#### Global methods

GenePix Pro also offers several global background subtraction methods. In a global method, a single value for the background is used for a whole array at each wavelength.

The global background values are based on local background regions, as explained above in “Local methods”. For example, when calculating a ‘mean of all



■ background pixels  
 ■ 2-pixel exclusion region  
 ■ feature pixels

Figure 1: defining background pixels

### **Consistency of slide preparation:**

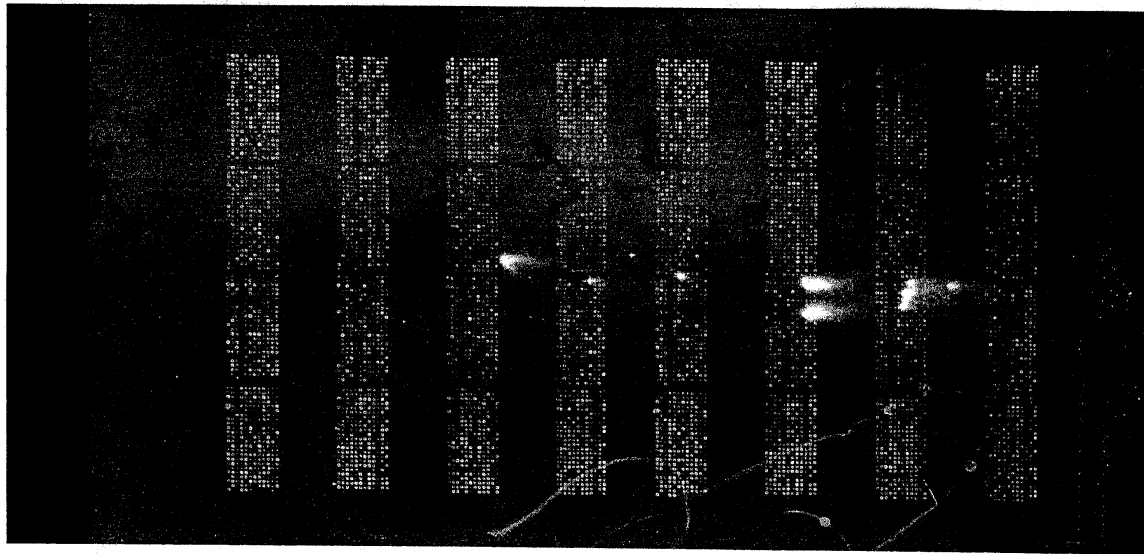
- inkjet-like technology; 8×4 print tips (broken/worn?)
- non-uniform spots across slide
- registration errors

### **Sample preparation: 'labeling efficiency:**

- equal volumes of expt & control cells?
- equal amounts of Cy3 (red) / Cy5 (green) ?
- equal levels of Cy3,Cy5 cell binding?

### **Instrumentation errors (laser scanner):**

- range of fluorescence levels, both channels
- scanning accuracy/precision, both channels
- sample degradation over time between scans (532nm, 635nm)



duke 01

	Stripe			
	1	2	3	4
	5	6	7	8
	..	..	..	..
Layer	..	..	..	..
	..	..	..	..
	25	26	27	28
	29	30	31	32

Each block has  $529 = 23 \text{ rows} \times 23 \text{ columns}$  of spots

Data Transformation:  $g$ -family of (lognormal) distributions:

$$(X - a)/b = (e^{gZ} - 1)/g$$

$Z \sim N(0, 1)$ ,  $a = \text{location}$ ,  $b = \text{scale}$ ,  $g = \text{skewness}$

$$Z = z(X) = g^{-1} \cdot \log[g \cdot (X - a)/b + 1]$$

Quantiles of Z and X transform consistently:

$$\begin{aligned} P\{ X \leq x_p \} &= P\{ Z \leq z_p \} = p \\ \Rightarrow z_p &= g^{-1} \cdot \log[g \cdot (x_p - a)/b + 1] \end{aligned}$$

Fitting  $a, b, g$ : Hoaglin (1985, EDTTS Ch 11):

$$\begin{aligned} x_p &= a + b(e^{gz_p} - 1)/g \\ x_{1-p} &= a + b(e^{-gz_p} - 1)/g \\ x_{0.5} &= a = x_M \text{ (median)} \\ g_p &= -(1/z_p) \cdot \log[(x_{1-p} - x_M)/(x_M - x_p)] \end{aligned}$$

Repeat for all 32 blocks

## *Background estimation*

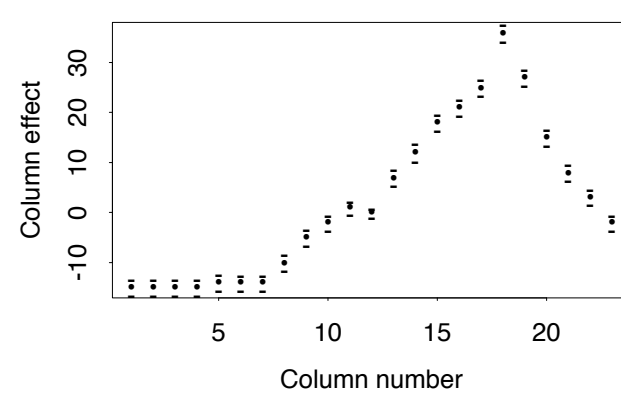
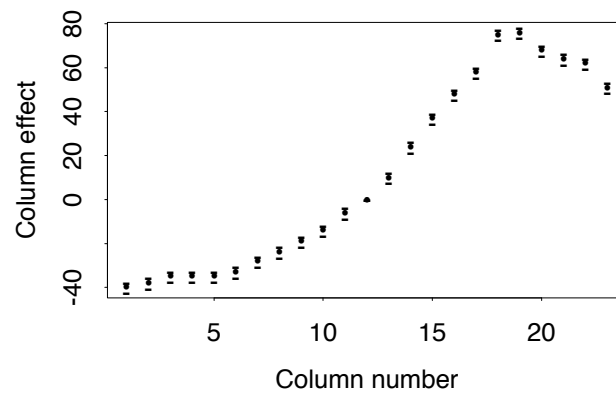
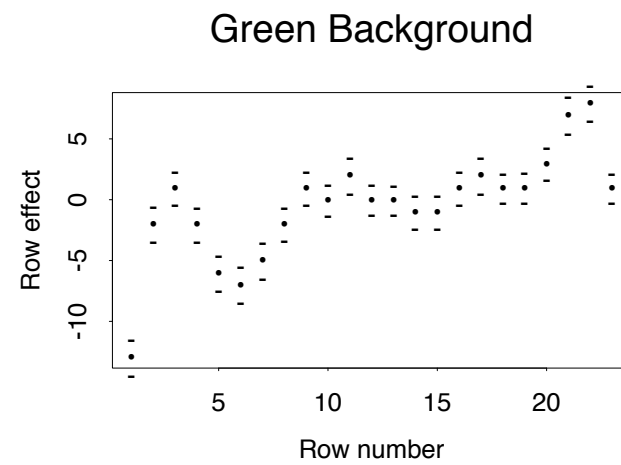
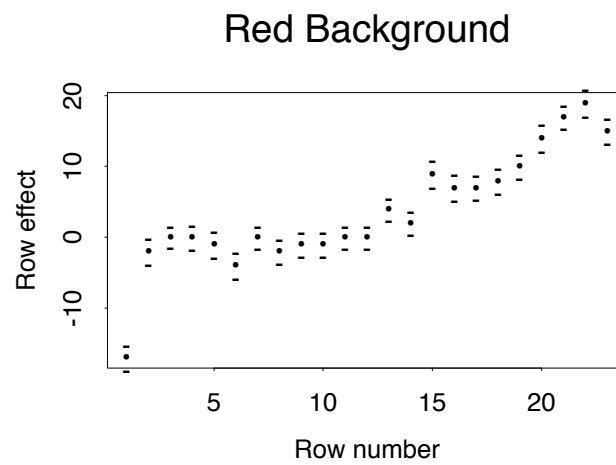
- Counts appear even in absence of spots  
(smearing, artifacts on slide, environment, ...)
- Adjustment: *foreground* – *background* (may be negative)
- Many background algorithms (Yang et al. 2002)

Block background medians: Two-way model

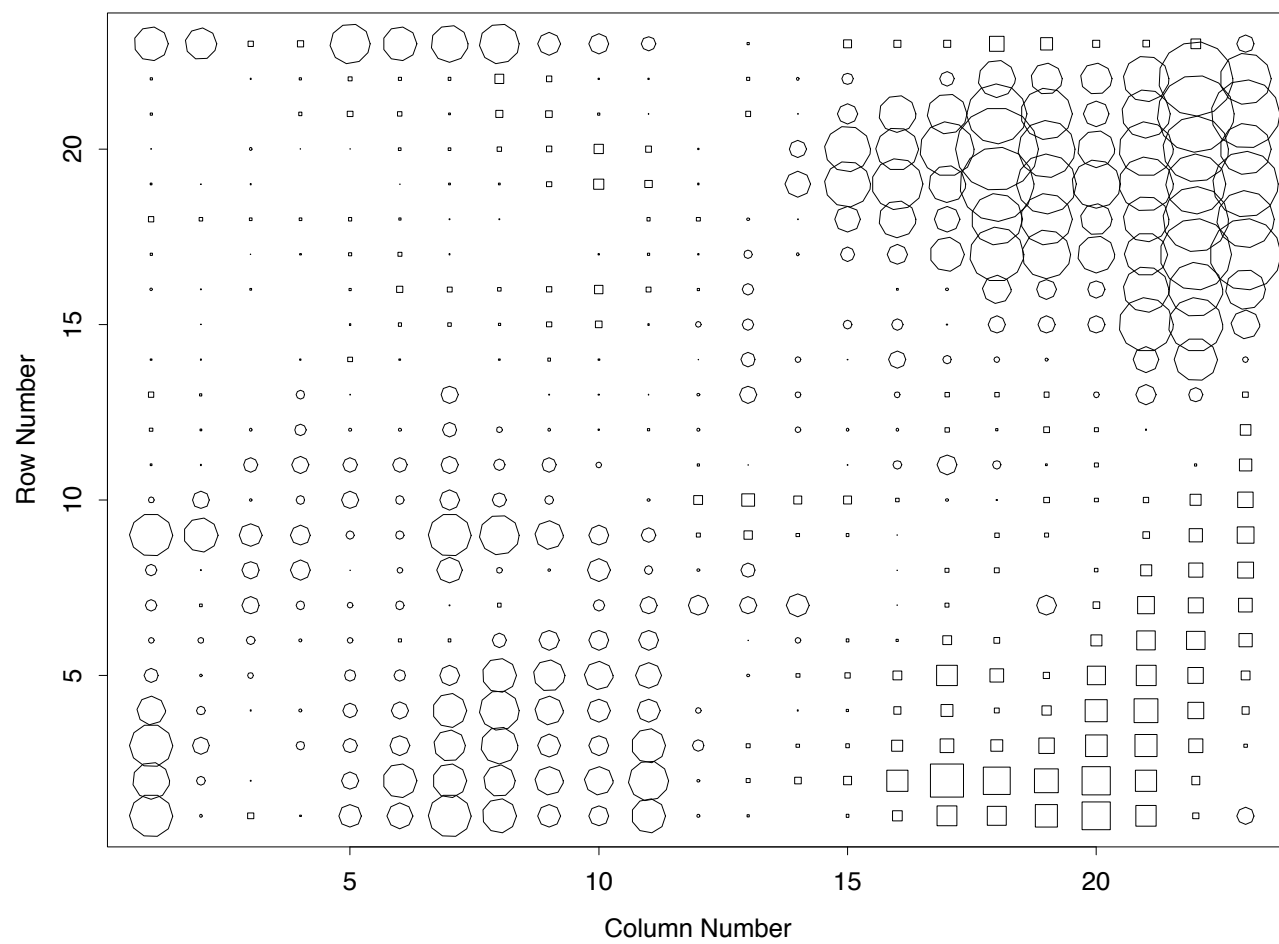
$$b_{ij} = m + row_i + col_j + res_{ij}$$

- Plot  $row_i$  vs  $i$  ( $i = 1, \dots, 23$  rows)
- Plot  $col_j$  vs  $j$  ( $j = 1, \dots, 23$  columns)
- “Plus-one” fit if residuals show structure (ODOFNA):

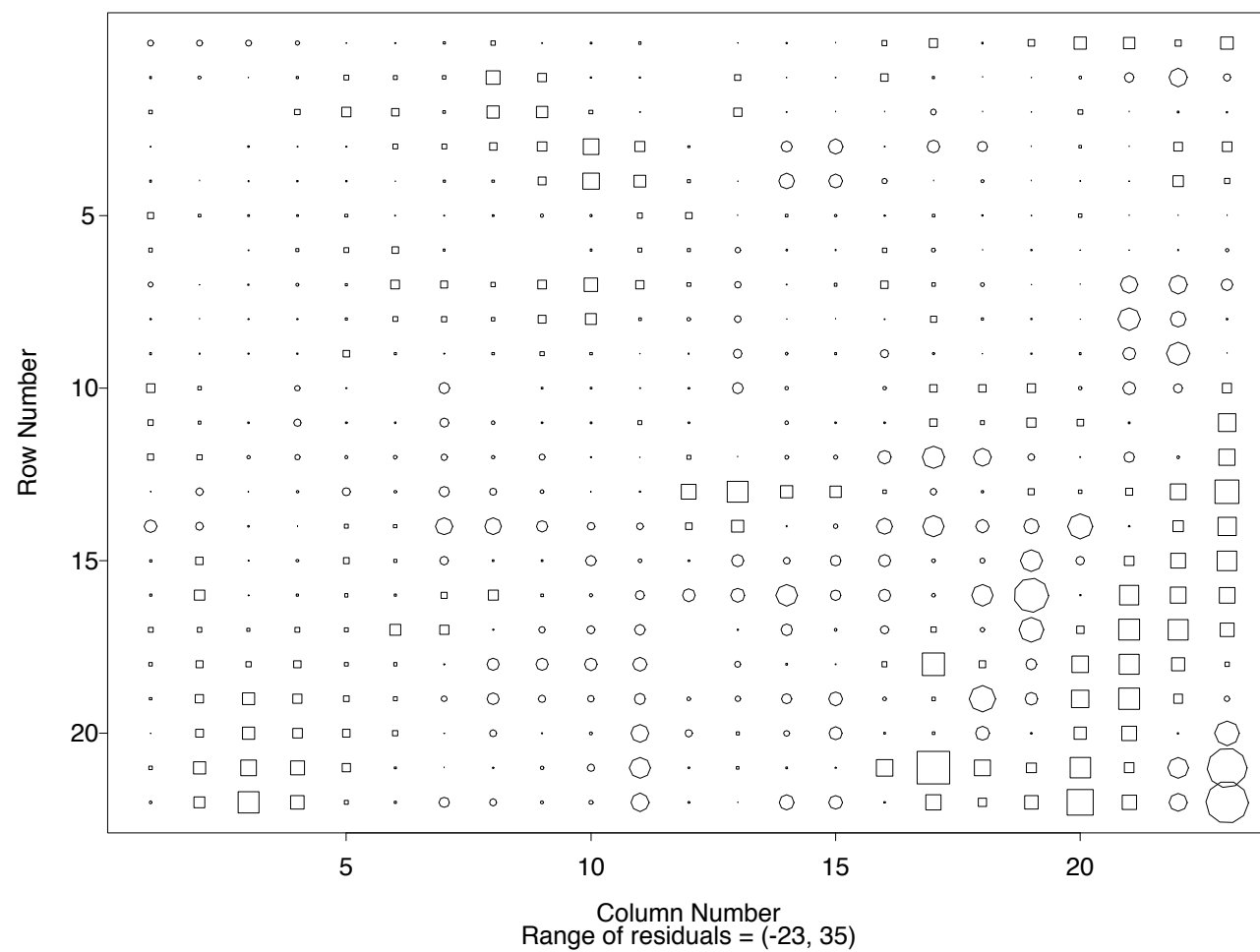
$$b_{ij} = m + row_i + col_j + T \cdot row_i \cdot col_j + res_{ij}$$







Coded residuals for Red background channel, Block 8



“ $m$ ” (common term) for each block:

8 “layers”  $\times$  4 “stripes” = 32 blocks

Fit two-way model to block terms:

$$m_{ls} = M + layer_l + stripe_s + res_{ls}$$

- Spatial effects on slide
- More stable estimate of background
- Subtract *fitted* background from foreground (fewer negatives)
- Transform adjusted foreground values  $\Rightarrow (Z_R, Z_G)$
- $(Z_R, Z_G)$  now approximately bivariate Gaussian
- Relationships among  $a_R, b_R, g_R, a_G, b_G, g_G$

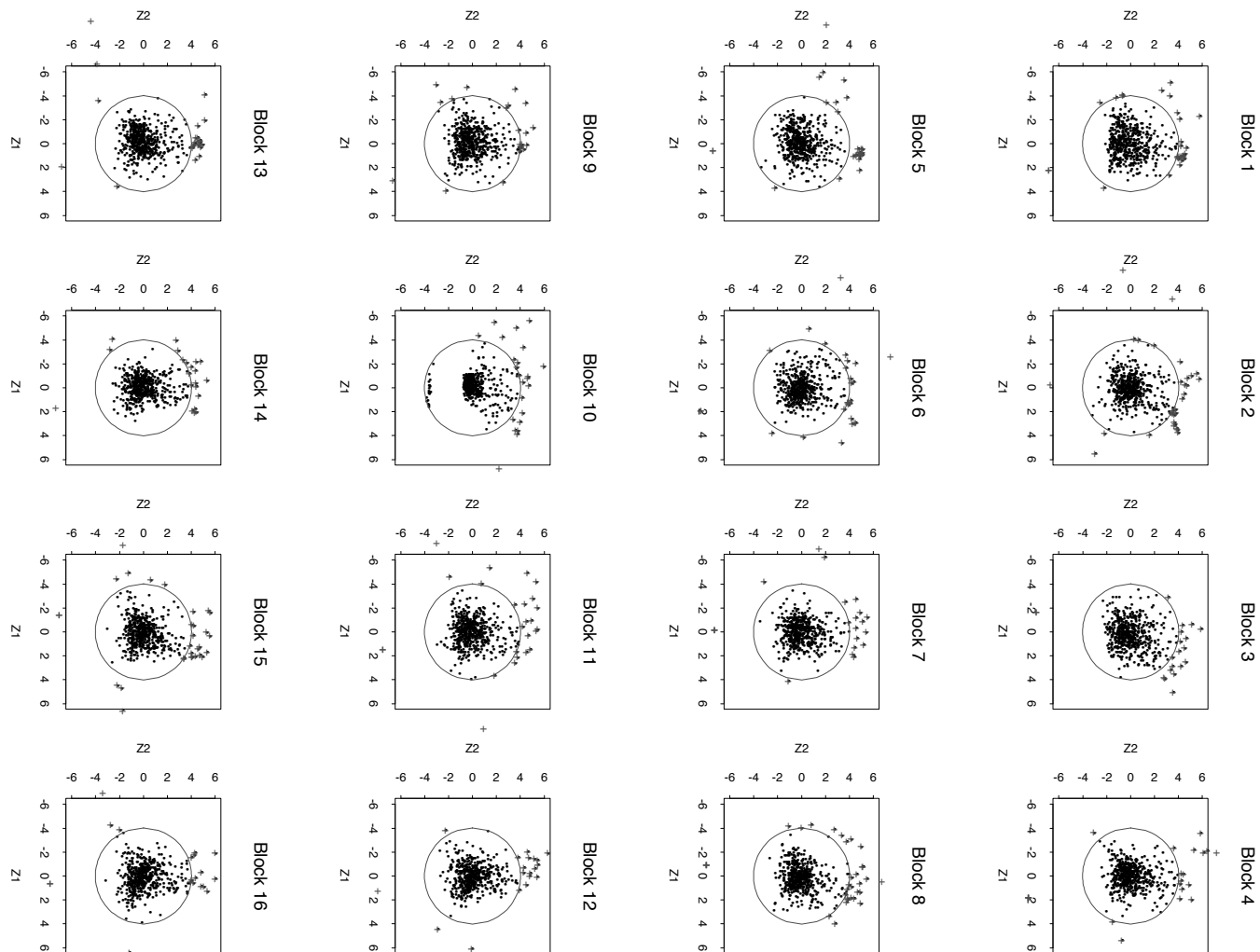
## Analysis steps

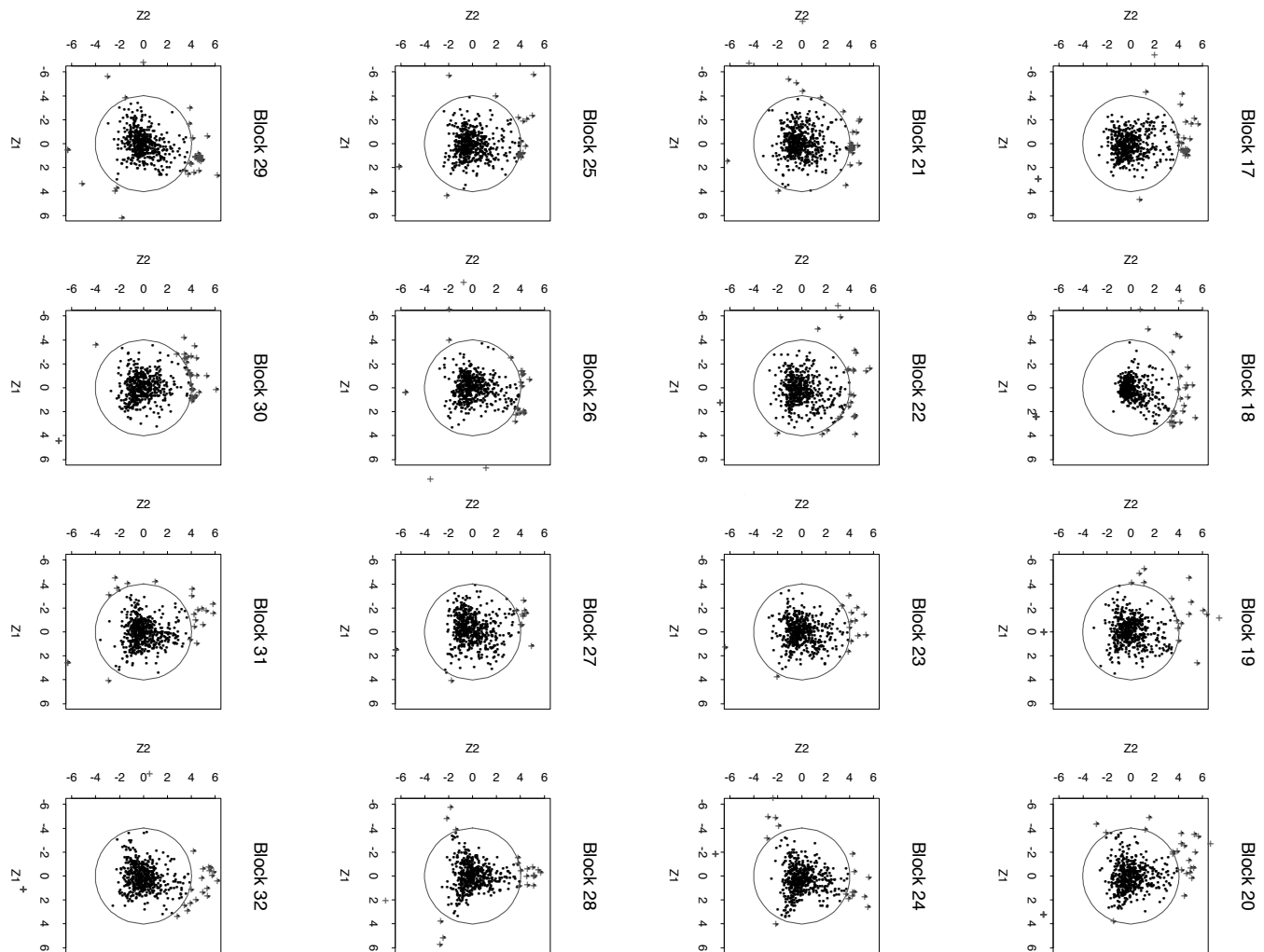
1. Median polish applied separately to background counts in each block (possibly with smoothing of the fitted row and column effects, and possibly with the extra term for non-additivity), yielding 32 sets of fitted background counts
2. Adjust foreground counts in each channel: subtract fitted background counts in Step 1 from reported foreground counts
3. Estimate for each block the parameters in  $g$  family to adjusted foreground counts obtained in Step 2
4. Transform adjusted foreground counts using  $g$ ,  $a$ ,  $b$ , to obtain  $\approx$  Gaussian quantities

$$Z_R = g_{kr}^{-1} \log[g_{kr}(R_{ijk}^* - a_{kr})/b_{kr} + 1]$$
$$Z_G = g_{kg}^{-1} \log[g_{kg}(G_{ijk}^* - a_{kg})/b_{kg} + 1]$$

5. Estimate correlation  $\hat{\rho}_k$  between  $Z_R$  and  $Z_G$  in each block:  
 $\text{cor}(Z_r, Z_g, \text{trim}=j/100)$
6. Calculate an approximate standard error for the difference  
 $Z_R - Z_G$  as  $[2(1 - \hat{\rho}_k)]^{1/2}$ , or
7.  $(Z_R^* = (Z_R - Z_G)/\sqrt{2(1 - \hat{\rho})}, Z_G^* = (Z_R + Z_G)/\sqrt{2(1 + \hat{\rho})}) \sim N_2(0, I)$
8. Weighted average  $(Z_R^*, Z_G^*)$  from several experiments
9. Denote as “significant” those  $(Z_R^*, Z_G^*)$  that fall outside a circle of radius 3 SEs

Results on these data: About 178 genes “significant”





## 4. Internet Traffic Data

Collected from anonymous surveillance machines outside “firewall” to monitor incoming/outgoing traffic [Marchette 2001, *Computer Intrusion Detection and Network Monitoring*]

- All internet communications are transmitted via packets
- Fundamental unit of information is **packet**
- Packet consists of data and headers that control the communications via IP, TCP
- **Flow** = exchange of packets between source-destination pair
- **Connections** = collections of flows (these data)
- After much pre-processing, data file has summary statistics on size/duration of **flows** (millions per hour)



#### 4. Sample Data Set From George Mason Univ.

length	SIP	DIP	DPort	SPort	Npkt	Nbyte
0.23	4367	54985	443	1631	9	3211
0.27	18146	9675	3921	25	15	49
0.04	18208	28256	1255	80	6	373
1389.10	24159	17171	23	1288	845	5906
373.99	60315	37727	2073	80	1759	834778
0.13	28256	18208	80	1256	10	816
1498.11	25699	4837	9593	80	65803	35661821
0.04	18208	28256	1251	80	5	373
122.38	54985	4179	1298	443	99	85559

## TCP

- Instructions for delivering/sequencing packets coming from one machine, destined for another
- Data passed through “ports” (logical, rather than physical, location; identifies connection through which data are passed between machines)
- $2^{16} = 65,536$  ports per machine
  - $2^{10}$  Ports 0–1023: “well-known ports”
  - Registered Ports 1024–49151 (e.g., 2049 for Sun’s `nfs`)
  - $2^{14}$  Ports 49152–65536: dynamic/private ports
- Unprotected ports are prime candidates for intrusion, so monitor amount of traffic in/out of ports

- Among the  $2^{10}$  “well-known” Ports 0–1023:

21	ftp	file transfer protocol
22	ssh/scp	secure shell/copy
23	telnet	network connection
25	smtp	mail transmission protocol
80	http	conventional web port (also 8080)
110	pop3	pop3 mail
443	https	secure web encryption
554	rtsp	real-time streaming video/audio

- DPort = **Destination Port**; SPort = **Source Port**

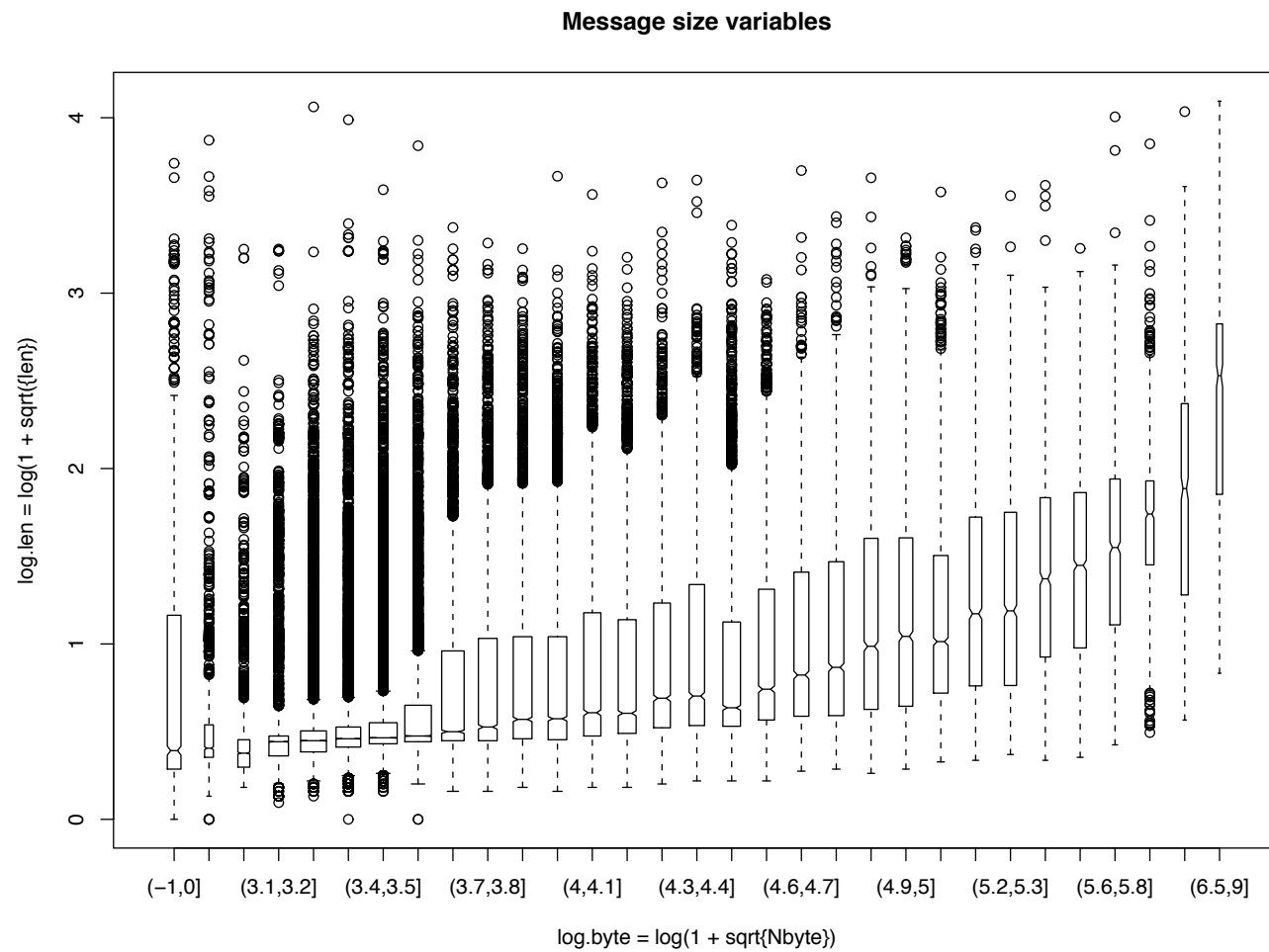
## Message sizes

Three (correlated) measures of “size” of session:

- `duration` or `len` = duration (length) in seconds
- `Nbyte` = Number of bytes
- `Npkt` = Number of packets
- Highly skewed distributions: use log transformation  
 $f(x) = \log(1 + \sqrt{x})$ : `log.len`, `log.byte`, `log.pkt`  
( $\log(x)$  spreads many small  $x$ 's too much)
- Potentially suspicious:  
Few packets, each with many bytes  
Many packets, each with very few bytes

Summary statistics over 135,605 records (1 hour in 2002)

	len	SIP	DIP	DPort	SPort	Npkt	Nbyte
min	0.0	259	259	20	20	2	0
10%	0.2	4930	4024	80	1187	9	568
25%	0.3	9765	8705	80	1369	10	860
med	0.6	20258	25164	80	1849	12	1832
75%	3.8	41282	45900	80	3681	21	7697
90%	21.5	62754	58202	80	10000	45	25161
max	3482.5	65276	65262	10000	10000	65803	35mil
#!	9101	2504	5139	380	6742	1056	29876



## 5. ‘Visual Analytics’: Exploratory Plots

Detecting “exotic” requires characterizing “typical”

- Size variables are highly skewed (already seen)
- Boxplots suggest excessive numbers of “outliers”
- Need better display of distribution

Letter value plots

## Letter value Displays

Estimate quantiles corresponding to tail areas  $2^{-k}$ :

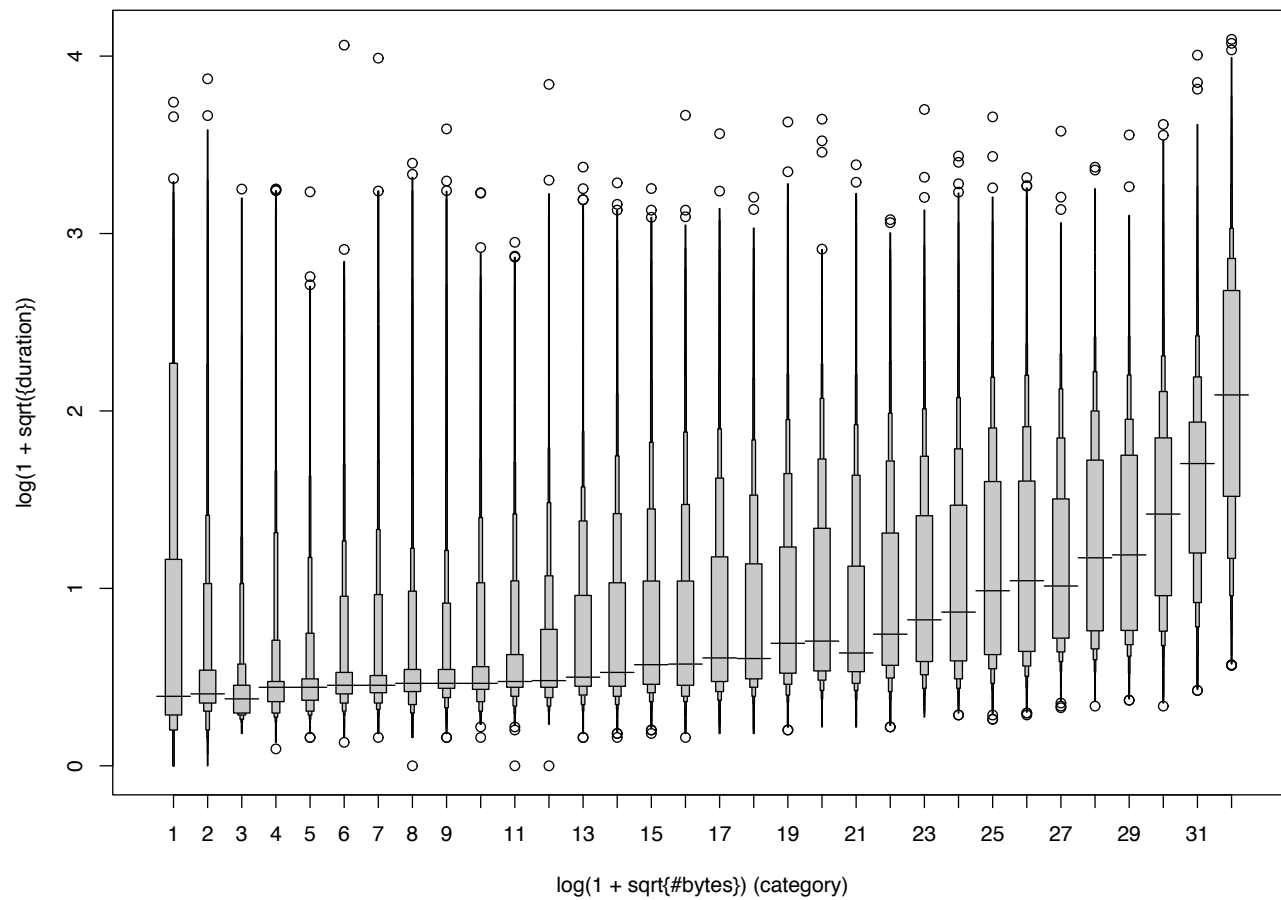
- Median ( $2^{-1}$ ): depth =  $d_M = (1 + n)/2$
- Fourths ( $2^{-2}$ ): depth =  $d_F = (1 + \lfloor d_M \rfloor)/2$
- Eighths ( $2^{-3}$ ): depth =  $d_E = (1 + \lfloor d_F \rfloor)/2$
- Etc. until extreme (depth = 1)
- When  $d_L$  is a half-integer, average two adjacent order statistics
- Actual tail area is closer to  $(d_L - \frac{1}{3})/(n + \frac{1}{3})$  [*URED*A §2G]
- Asymptotic correlation between adjacent LVs  $\approx \sqrt{\frac{1}{2}} = 0.707$



## Letter value box plots

- Small data sets: Limited information about tails
- Boxplots show fourths, *extent* of data beyond fourths
- Large data set: Tail quantiles more reliable
- $\Rightarrow$  Extend boxplots to include more letter values beyond median, fourths
- “Stopping Rules”: how many LVs to show?
- How to display letter values?
- Which observations are labeled as ‘outliers’?
- **Plot still shows only actual data values**

Message Size and Length, 135,605 Sessions



### More exploratory plots

- Preponderance of relatively short sessions (not shown)
- Number of active sessions in 120 successive 30-second non-overlapping intervals (mean 923, SD 140,  $\approx 3\sigma$  limit 1343, max 1299 = expected max of 120  $N(923, 140^2)$  variates)
- Plot of `log.len` vs *Session Start Time* should be relatively uninteresting

## Evolutionary Displays for Internet data

- Exploratory plots are useful for modeling activity sessions
- Adapt evolutionary plots to summarized data
- Can we summarize as quickly as the data arrive?
- Fast algorithms can do little more than compare and add (linear operations)
- Robust methods usually rely on medians and sorts
- Fast + robust  $\Rightarrow$  compare and (keep or discard), follow by linear operations

## Waterfall Diagrams (Wegman and Marchette 2003)

- “Streaming plot”: Plot a point at location  $(s, t)$ , where  $t$  (time) = session start time (starts at 0, continues upward),  $s$  is a source IP (SIP) or source port (SPort)
- SIP: 4837 (occurs 4,754 times), 13525 (occurs 4,448 times), 65246 (occurs 12,150 times)
- SPort: Trends *across* plot may indicate scanning SPorts
- Useful for monitoring attempted access: For a given session (exchange of packets), initial port may be assigned arbitrarily; subsequent ones assigned by incrementing pattern characteristic of operating system. Attacker can tell from pattern of SPort increments about operating system

## Skyline plots

- “Streaming plot”: Plot access of DPorts or SPorts
- Recall: 353 of the 380 DPorts (92.9%) occur  $< 5$  times; the 3 most frequent DPorts are expected (`http`, `https`, `smtp`)
- Plot X when DPort is accessed (apart from “well-known” ports 0–1023); red X if count  $> 10$
- Likewise for SPort but higher “control limit” 12% (4%) of the 6742 unique SPorts occur  $> 50$  (100) times
- Building a “skyline”

## 5. High-Energy Physics (HEP)

- HEP experiments: Colliding beams of particles (MeV, GeV)
- Each collision (event)  $\Rightarrow$  more particles (products)
- Huge detector (“bandshell” of wires): detects final products
- Thousands of computer programs:
  - reconstruct particle tracks; connect particles w/tracks
  - identify (?) particles (ex:  $+$ , mass  $0.14\text{GeV}/c^2 \Rightarrow \pi^+$ )
  - estimate particle lifetime, momentum, mass

### Goals:

- **What happened?** (decay type)
- **How?** (Parameter of model for specific decay)

## Particles and Anti-particles: Classes and Properties

Fundamental particles: Fermions = Quarks  $\cup$  Leptons

- Quarks: 6 *flavors*

$u, c, t$ : charge  $\frac{2}{3}$

$d, s, b$ : charge  $-\frac{1}{3}$

- Leptons: 6 *flavors*

$e, \mu, \tau$ : charge  $-1$

$\nu_e, \nu_\mu, \nu_\tau$ : charge 0

Quarks combine to form all other particles:

proton  $uud$ ; anti-proton  $\bar{u}\bar{u}\bar{d}$ ; neutron  $uud$  ( $\sim 120$  *baryons*  $qqq$ )

pion  $\pi^+ = u\bar{d}$ ; kaon  $K^- = s\bar{u}$ ;  $B^+ = b\bar{u}$ ;  $B^- = \bar{b}u$






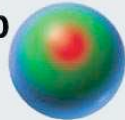
$B^0 = \bar{b}d$ ;  $\bar{B}^0 = b\bar{d}$  ( $\sim 140$  *mesons*  $q\bar{q}$ )









# PARTICLES OF MATTER:

## QUARKS






These particles make up protons, neutrons and a veritable zoo of lesser-known particles. They have never been observed in isolation.

<p>UP  <b>u</b></p> <p><b>Electric charge:</b> <math>+2/3</math>  <b>Mass:</b> 2 MeV            Constituent of ordinary matter; two up quarks, plus a down, make up a proton.</p>	<p>DOWN  <b>d</b></p> <p><b>Electric charge:</b> <math>-1/3</math>  <b>Mass:</b> 5 MeV            Constituent of ordinary matter; two down quarks, plus an up, compose a neutron.</p>
<p>CHARM  <b>c</b></p> <p><b>Electric charge:</b> <math>+2/3</math>  <b>Mass:</b> 1.25 GeV            Unstable heavier cousin of the up; constituent of the J/theta particle, which helped physicists develop the Standard Model.</p>	<p>STRANGE  <b>s</b></p> <p><b>Electric charge:</b> <math>-1/3</math>  <b>Mass:</b> 95 MeV            Unstable heavier cousin of the down; constituent of the much studied kaon particle.</p>
<p>TOP  <b>t</b></p> <p><b>Electric charge:</b> <math>+2/3</math>  <b>Mass:</b> 171 GeV            Heaviest known particle, comparable in mass to an atom of osmium. Very short-lived.</p>	<p>BOTTOM  <b>b</b></p> <p><b>Electric charge:</b> <math>-1/3</math>  <b>Mass:</b> 4.2 GeV            Unstable and still heavier copy of the down; constituent of the much studied B-meson particle.</p>

# PARTICLES OF MATTER:

LEPTONS	
These particles are immune to the strong force and are observed as isolated individuals. Each neutrino shown here is actually a mixture of neutrino species, each of which has a definite mass of no more than a few eV.	
<p><b>ELECTRON NEUTRINO</b> <math>\nu_e</math></p>  <p><b>Electric charge: 0</b> Immune to both electromagnetism and the strong force, it barely interacts at all but is essential to radioactivity.</p>	<p><b>ELECTRON</b> <math>e</math></p>  <p><b>Electric charge: -1</b> <b>Mass: 0.511 MeV</b> The lightest charged particle, familiar as the carrier of electric currents and the particles orbiting atomic nuclei.</p>
<p><b>MUON NEUTRINO</b> <math>\nu_\mu</math></p>  <p><b>Electric charge: 0</b> Appears in weak reactions involving the muon.</p>	<p><b>MUON</b> <math>\mu</math></p>  <p><b>Electric charge: -1</b> <b>Mass: 106 MeV</b> A heavier version of the electron, with a lifetime of 2.2 microseconds; discovered as a component of cosmic-ray showers.</p>
<p><b>TAU NEUTRINO</b> <math>\nu_\tau</math></p>  <p><b>Electric charge: 0</b> Appears in weak reactions involving the tau lepton.</p>	<p><b>TAU</b> <math>\tau</math></p>  <p><b>Electric charge: -1</b> <b>Mass: 1.78 GeV</b> Another unstable and still heavier version of the electron, with a lifetime of 0.3 picosecond.</p>

**PARTICLES OF FORCE:**

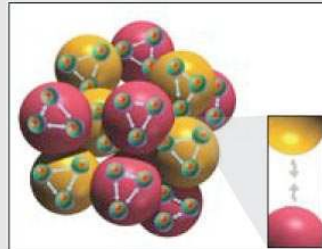
<b>BOSONS</b> At the quantum level, each force of nature is transmitted by a dedicated particle or set of particles.	
<b>PHOTON</b>  $\gamma$  <b>Electric charge: 0</b> <b>Mass: 0</b> Carrier of electromagnetism, the quantum of light acts on electrically charged particles. It acts over unlimited distances.	<b>Z BOSON</b>  <b>Z</b>  <b>Electric charge: 0</b> <b>Mass: 91 MeV</b> Mediator of weak reactions that do not change the identity of particles. Its range is only about $10^{-18}$ meter.
<b><math>W^+/W^-</math> BOSONS</b>  <b>W</b>  <b>Electric charge: +1 or -1</b> <b>Mass: 80.4 GeV</b> Mediators of weak reactions that change particle flavor and charge. Their range is only about $10^{-18}$ meter.	<b>GLUONS</b>  <b>g</b>  <b>Electric charge: 0</b> <b>Mass: 0</b> Eight species of gluons carry the strong interaction, acting on quarks and on other gluons. They do not feel electromagnetic or weak interactions.
<b>HIGGS</b>  <b>H</b> (not yet observed)  <b>Electric charge: 0</b> <b>Mass: Expected below 1 TeV, most likely between 113 and 192 GeV.</b> Believed to endow $W$ and $Z$ bosons, quarks and leptons with mass.	

### HOW THE FORCES ACT

An interaction among several colliding particles can change their energy, momentum or type. An interaction can even cause a single particle in isolation to decay spontaneously.

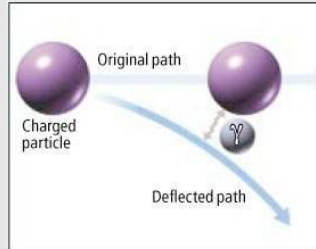
#### ELECTROMAGNETIC INTERACTION

The electromagnetic interaction acts on charged particles, leaving the particles unchanged. It causes like-charged particles to repel.



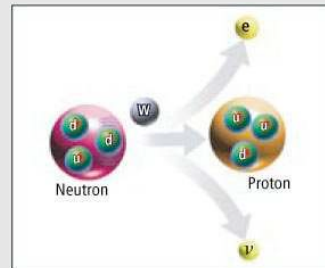
#### STRONG INTERACTION

The strong force acts on quarks and gluons. It binds them together to form protons, neutrons and more. Indirectly, it also binds protons and neutrons into atomic nuclei.



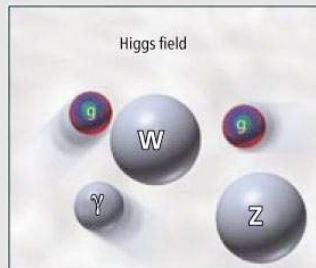
#### WEAK INTERACTION

The weak interaction acts on quarks and leptons. Its best-known effect is to transmute a down quark into an up quark, which in turn causes a neutron to become a proton plus an electron and a neutrino.



#### HIGGS INTERACTION

The Higgs field (*gray background*) is thought to fill space like a fluid, impeding the W and Z bosons and thereby limiting the range of weak interactions. The Higgs also interacts with quarks and leptons, endowing them with mass.



Examples from chart of Standard Model (<http://CPEPweb.org>)

1. neutron ( $udd$ )  $\beta$ -decay to a proton ( $uud$ ), an electron ( $e^-$ ), and an antineutrino ( $\bar{\nu}_e$ )
2. electron-positron ( $e^+e^-$ ) collision  $\Rightarrow$  meson pair  $B^0\bar{B}^0$  via a virtual  $\gamma$  photon or virtual  $Z$  boson
3. proton pair (pp) collision  $\Rightarrow$  many hadrons (baryons  $qqq$  or  $\bar{q}\bar{q}\bar{q}$  or mesons  $q\bar{q}$ ) and bosons (force carriers)

Masses, charges, spins of  $q$ 's (u,d,t,b,c,s), plus theory of behavior (Standard Model), determine masses, charges, spins of  $\sim 120$  currently known baryons ( $\sim 140$  mesons)

The Standard Model summarizes the current knowledge in Particle Physics. It is the quantum theory that includes the theory of strong interactions (quantum chromodynamics or QCD) and the unified theory of weak and electromagnetic interactions (electroweak). Gravit is included in the chart because it is one of the fundamental interactions even though not part of the "Standard Model".

The Standard Model summarizes the current knowledge in Particle Physics. It is the quantum theory that includes the theory of strong interactions (quantum chromodynamics or QCD) and the unified theory of weak and electromagnetic interactions (electroweak). Gravity is included on the chart because it is one of the fundamental interactions even though not part of the "Standard Model".

matter constituents  
spin = 1/2, 3/2, 5/2, ...

**Example 10** The angular momentum of particles, ions is given in table of 10, which is the maximum limit of angular momentum, where  $\hbar = 1.05 \times 10^{-34}$  J.s and  $\pi = 3.14$ .



spin = 0, 1, 2, ...

spin = 0, 1, 2, ...

ally charged particles resistant to exchanging photons, in strong interactions (color-charged) they interact by exchanging gluons. Leptons, photons, and  $W$  and  $Z$  bosons have no strong interactions and are color chargeless.

**Quarks Confined in Mesons and Baryons**

One cannot isolate quarks and gluons; they are confined in color-neutral particles called hadrons. The most common hadrons are mesons and baryons. Mesons consist of a quark and an anti-quark combination. As color-charged particles (quarks and gluons) move apart, the energy in the color field lines between them increases. This energy eventually is converted into mass and creates new quark-antiquark pairs. These quarks and antiquarks then combine with the original ones and the particles cease to interact. Thus, quarks of hadrons have been identified as **mesons** and **baryons** only.

[illegible]

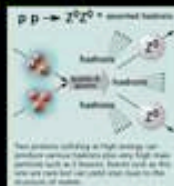
Variable	Factor	Sum of Squares	Degrees of Freedom	Mean Square	Value of $\eta^2$
$\eta^2$	gender	1.00	1	1.000	0.000
$K^+$	gender	1.00	1	1.000	0.000
$\mu^+$	gender	1.00	1	1.000	0.000
$\eta^0$	gender	1.00	1	1.000	0.000
$\eta^-$	gender	1.00	1	1.000	0.000

Mutter and Antismeyer

For every lattice  $\Lambda$  there is a corresponding antiparticle  $\Lambda^*$ , denoted by a bar over the particle symbol (lattice  $\bar{\Lambda}$  or  $\Lambda^*$  charge is opposite). Particle and antiparticle have identical mass and spin but opposite charges. Some obviously neutral lattices (e.g.,  $\mathbb{Z}^2$ ,  $\mathbb{A}_1$ , and  $\mathbb{A}_2 + \mathbb{A}_2'$ , but not  $\mathbb{E}_7 + \mathbb{A}_1$ ) are their own antiparticles.

**Figures**

These diagrams are an artistic conception of physical processes. They are not exact, and have no meaningful scale. Lines shaded gray represent the flow of photons in the given field, and red lines the given matter.

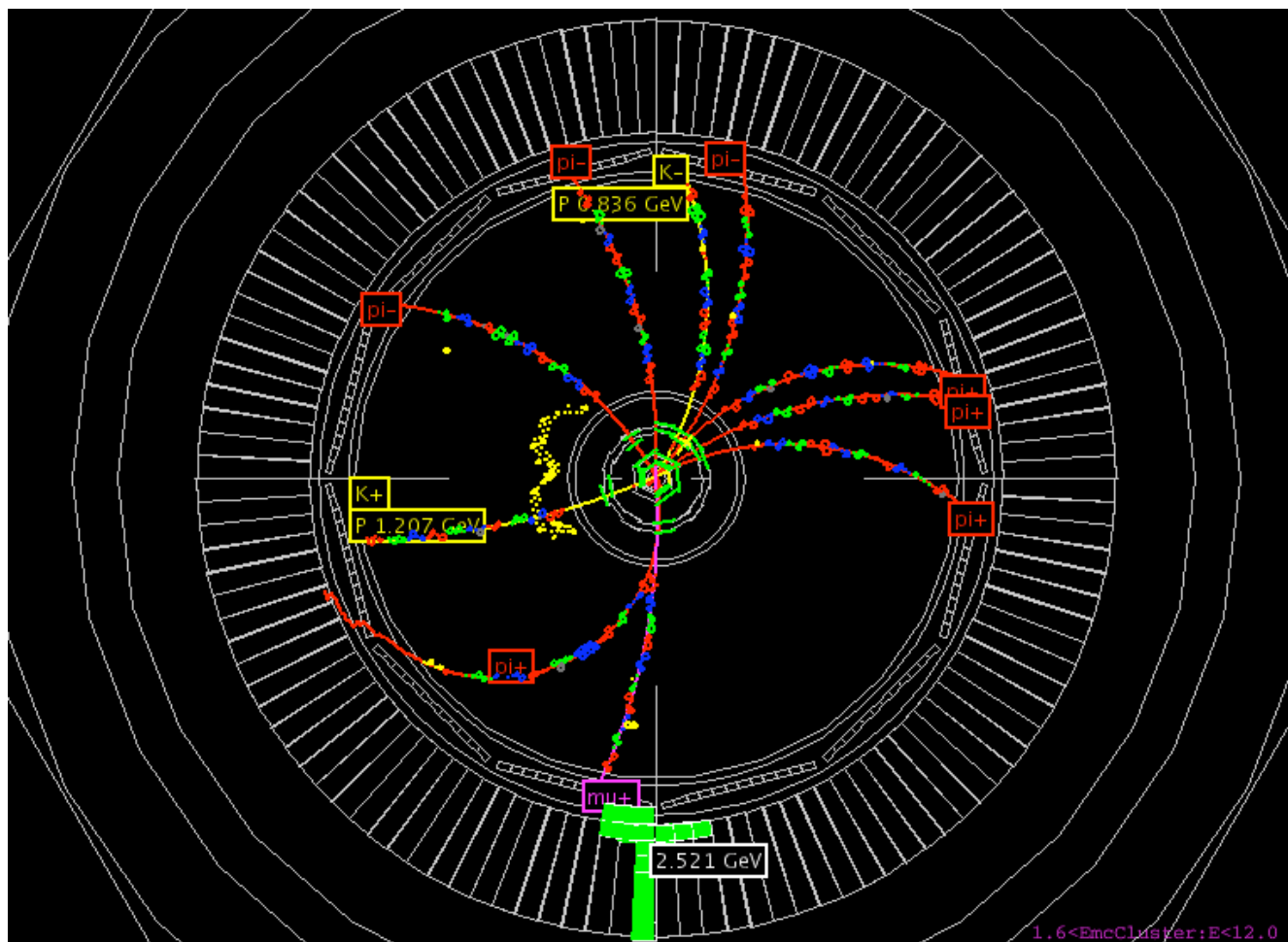


**The Particle Advantage**  
 Visit the journal online today! Features The Particle Advantage at  
<http://nsls.slac.stanford.edu/journal/online>

**Acknowledgments**  
This study has been made possible by the generous support of  
U.S. Department of Energy  
Lawrence Berkeley Laboratory  
Stanford University Academic Center  
Lawrence Livermore National Laboratory  
Lawrence Radiation Laboratory, University of California, Berkeley

©1994-1998 Communication Media Education Project (CME) is a nonprofit organization of teachers, physicians, and educators. Send mail to: CME, Attn: Ms. Jo Ann, Gateway Research National Laboratory, Berkeley, CA 94702. For information on CME's free materials, methods, research activities, and workshops, see <http://edg.berkeley.edu/cme.html>

- Hundreds of thousands of possible decays (events)
- $\sim 1$  event per 10 nanoseconds
- Data collection for 1 **day** would fill **200 DVDs**
- Massive filtering steps (“triggers”) to discard data from 99+% of events whose mechanisms are well understood
- SLAC saves data from  $\sim 100$  **events per second**
- Sensor resolution: may miss some particles from event
- Mis-ID: may connect particle tracks with wrong event
- Mis-reconstruct: right tracks, wrong reconstruction





SLAC collides beams of electrons, at energies designed to generate events whose products involve many  $b$ 's (*'b factory'*):

$$B\text{-mesons } B^0 = \bar{b}d, \bar{B}^0 = b\bar{d}$$

Example event sequence:

1. Energy from collision produces two quarks,  $b$  and  $\bar{b}$
2. Remaining energy from collision creates other  $q\bar{q}$  pairs ( $u\bar{u}$ ,  $d\bar{d}$ , ...), plus kinetic energy
3.  $q\bar{q} = u\bar{u} \Rightarrow$  charged B-mesons  $b\bar{u} \equiv B^+$ ,  $\bar{b}u \equiv B^-$  ( $\sim 50\%$ )  
 $q\bar{q} = d\bar{d} \Rightarrow$  neutral B-mesons  $\bar{b}d \equiv B^0$ ,  $b\bar{d} \equiv \bar{B}^0$  ( $\sim 50\%$ )
4. B-mesons then go on to decay in one of thousands of ways

Note: About 0.1% of collisions producing  $b - \bar{b}$  have insufficient energy to create new quarks, so  $b\bar{b}$  stays together as an  $\nu$  particle

Example: “target” decay mode of interest:

1.  $b$ ’s are produced ( $\sim 20\%$  of saved events)
2. energy from  $b - \bar{b}$  converts to mass of  $d\bar{d}$  quark pair  
 $\Rightarrow B^0$  or  $\bar{B}^0$  (**neutral** B-mesons)
3. **Either**  $B^0$  **or**  $\bar{B}^0 \rightarrow \rho^+\rho^- \rightarrow \pi^0\pi^+ \cup \pi^0\pi^-$
4. Expect at least 6–10 tracks per event (4 pions)

### Questions

- How often do such events occur? ( $10^{-5}$ )
- How long do sub-particles live? (e.g.,  $10^{-13}$  sec)
- Did anything else occur? (histogram, Poisson counts)

- Guiding principle:  
Less ‘noise’  $\Rightarrow$  Better estimates of decay rates
- Use data to rule out likely “imposters”  
(#tracks, final  $E_f < \text{or } \approx E_0$ , etc.)
- Data:  $\Delta E, m_B, Tthr, m_{\rho^+}, m_{\rho^-}, H_1, H_2$   
— *assuming our target decay occurred!*
- Physicists use *simulations* to predict features of data from our target event (uni-/bivariate pdfs); discard events where data lie outside ‘cuts’ of ‘likely’ regions on each variable
- Some *real* ‘imposter’ data is generated from events at different energies; assume frequency of ‘imposter’ events is the same at energy where target decay occurs
- Likelihood ratio test on all possible “imposters” not practical
- But we might be able to reduce “background” by considering

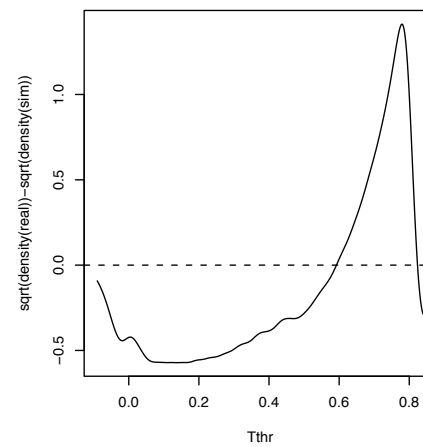
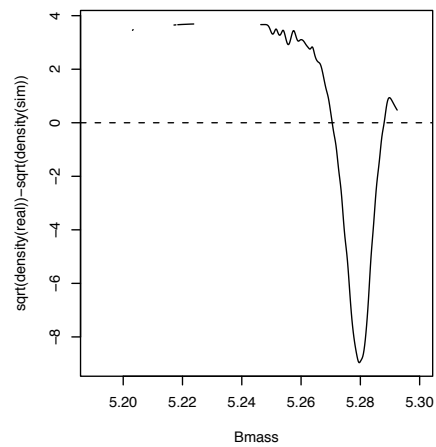
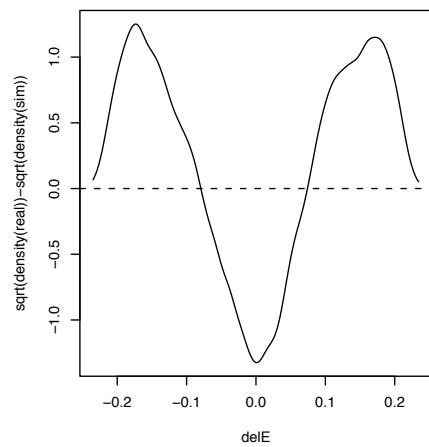
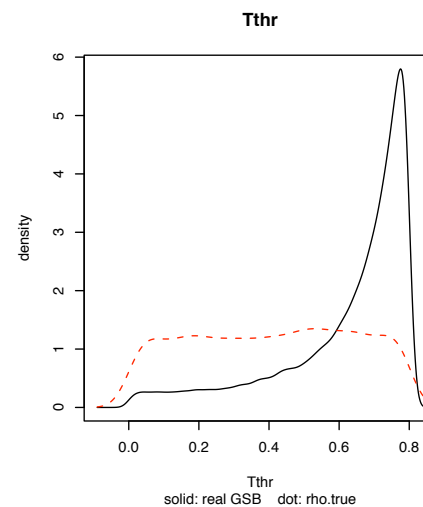
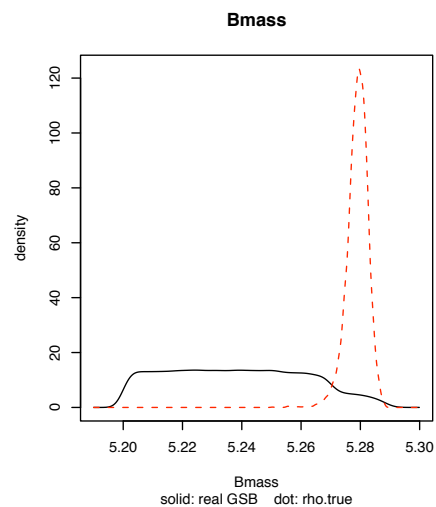
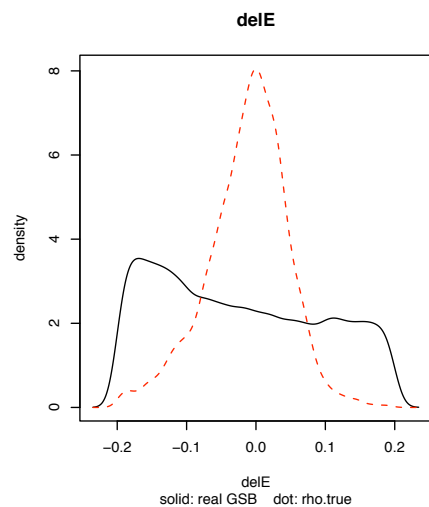
“likelihoods” from “top 20” most likely imposter decays

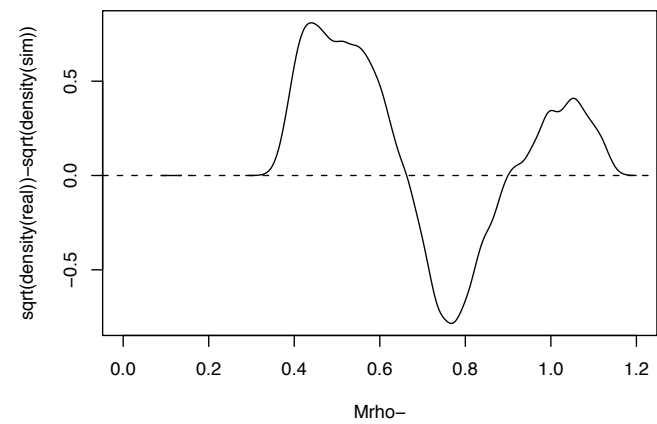
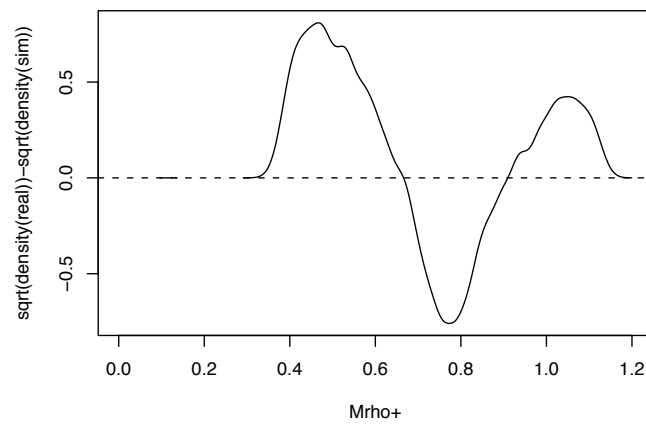
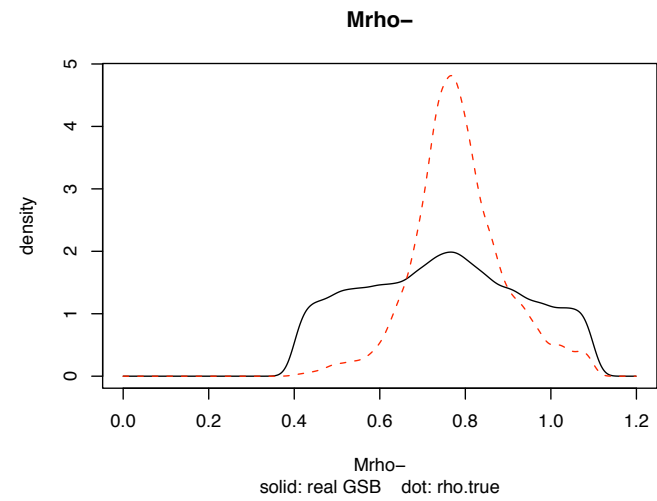
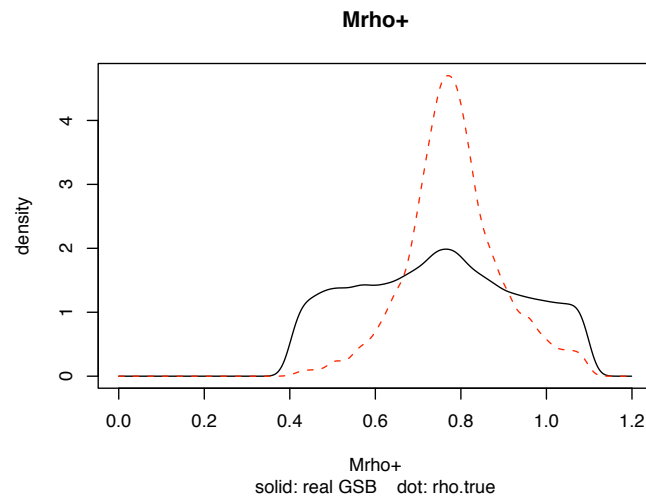
*Are ‘background’ events sufficiently different from ‘signal’ events?*

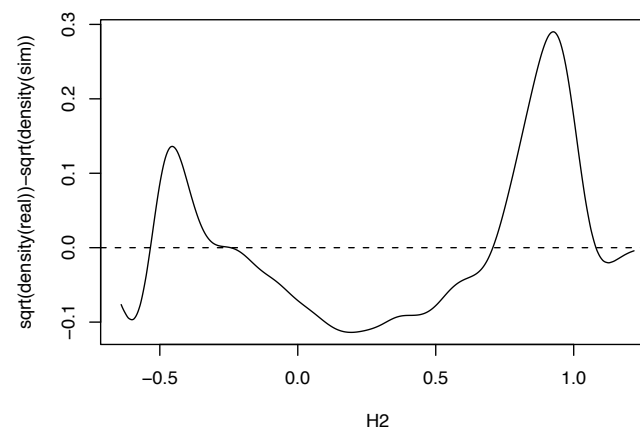
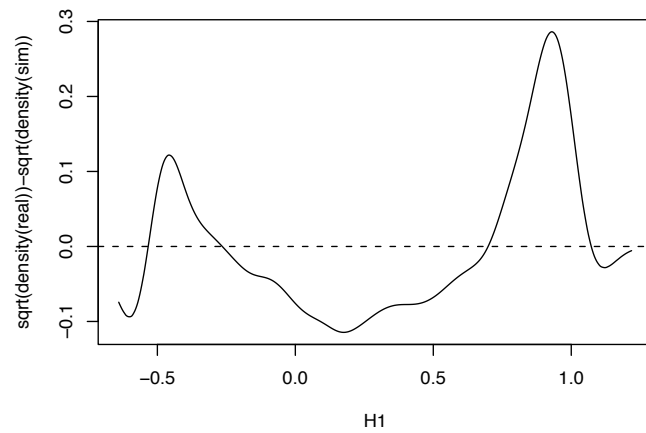
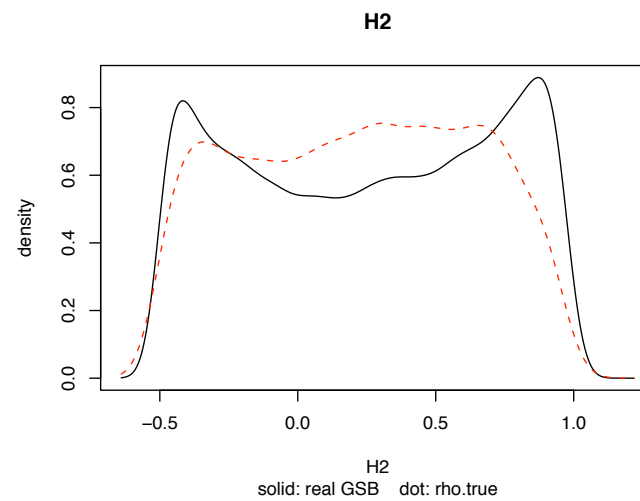
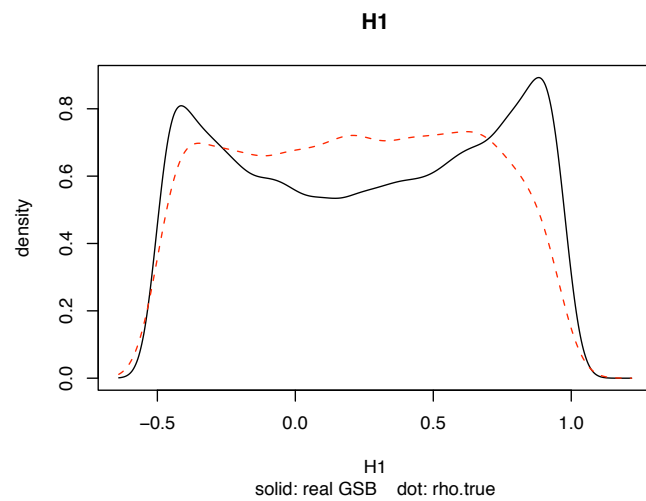
*How realistic are the simulated data sets?*

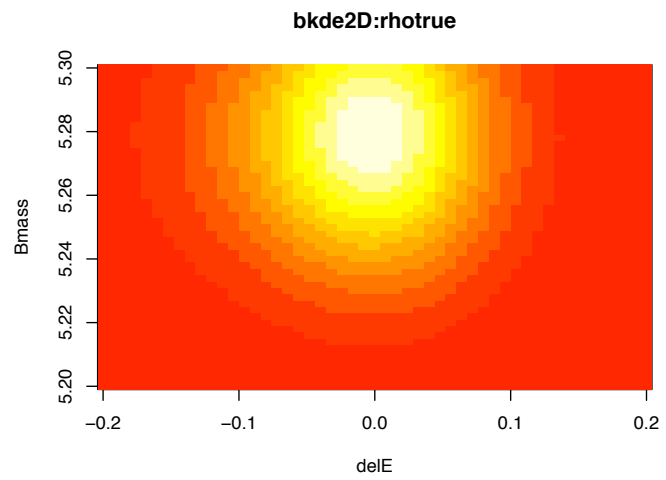
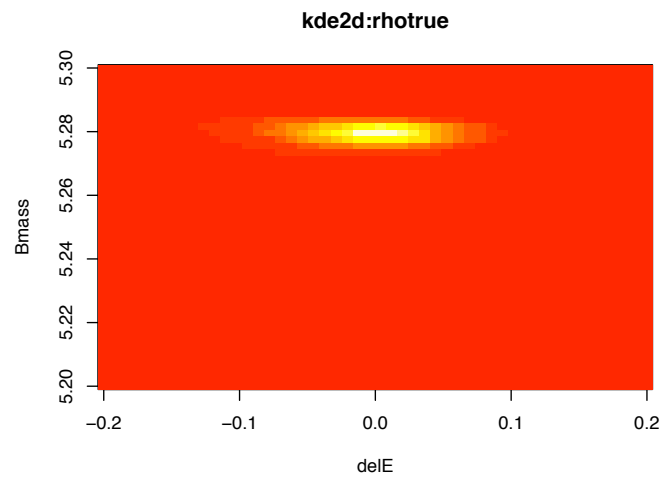
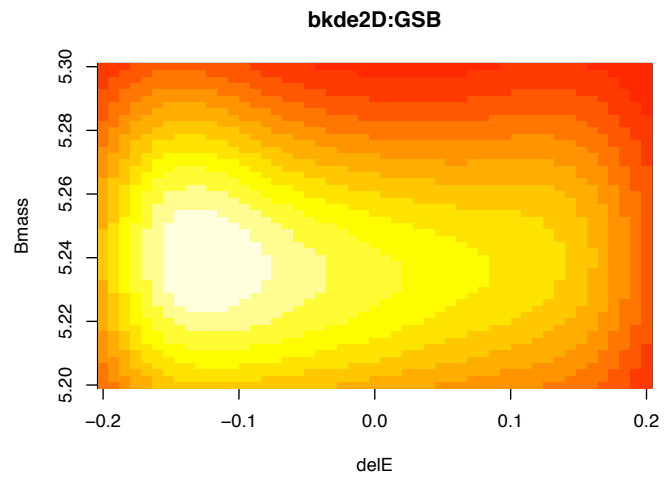
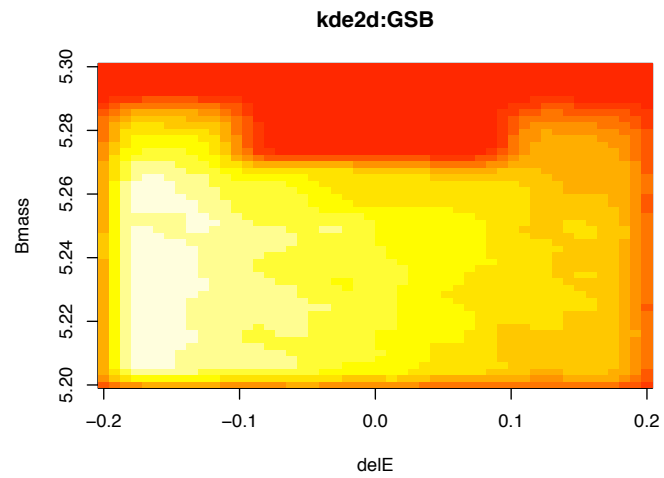
*Can we approximate the “likelihoods” of the “top 20” events?*

*How well do mixture models fit with non-Gaussian components,  
tiny signals?*

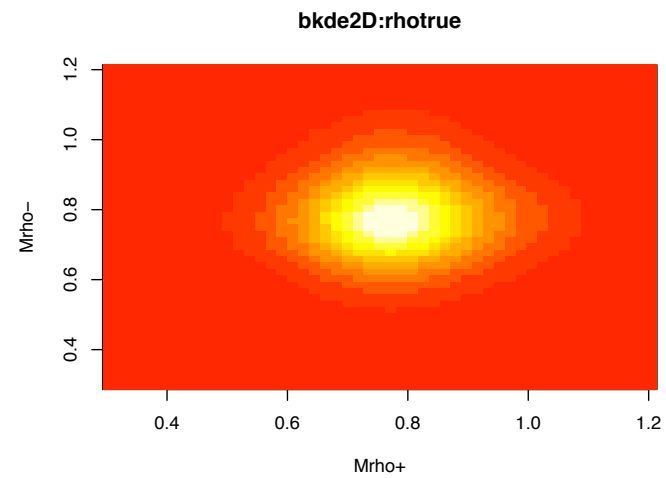
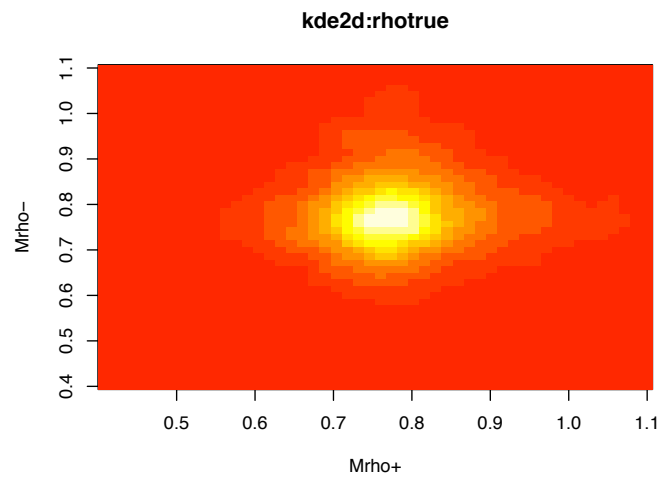
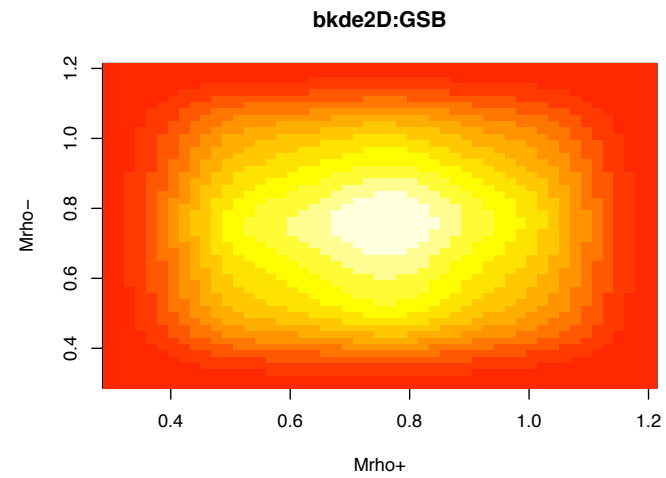
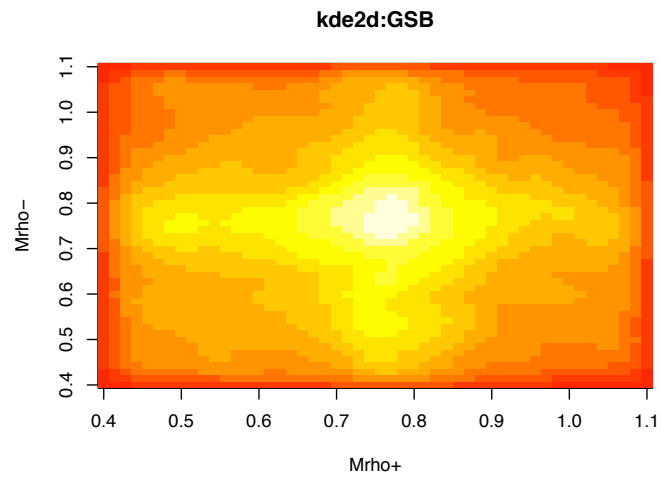


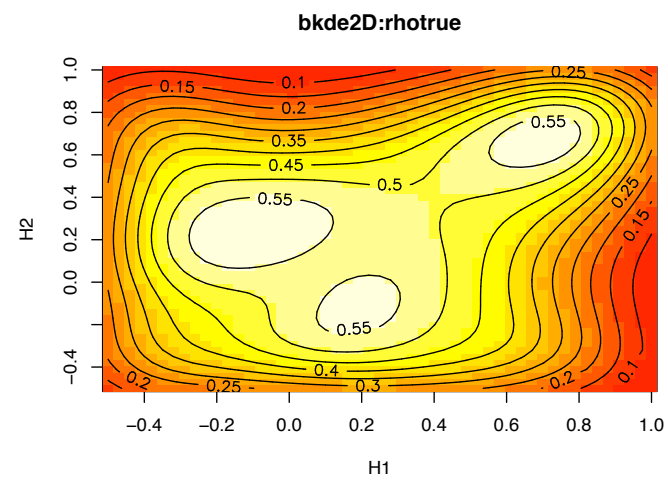
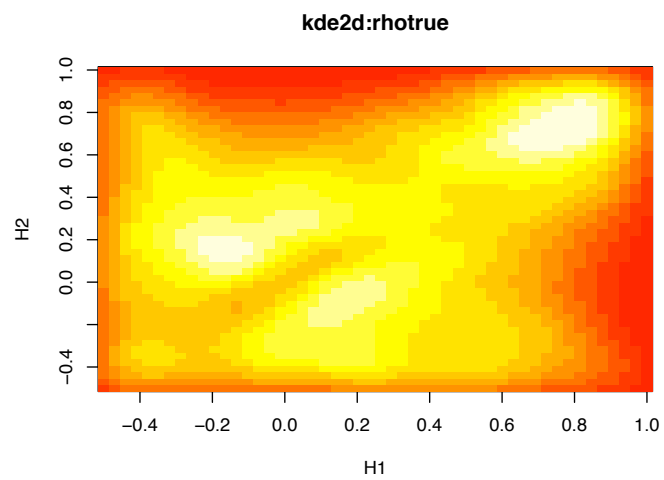
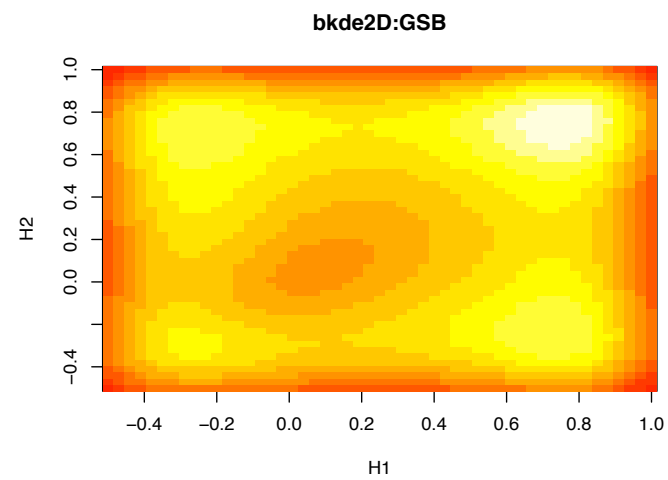
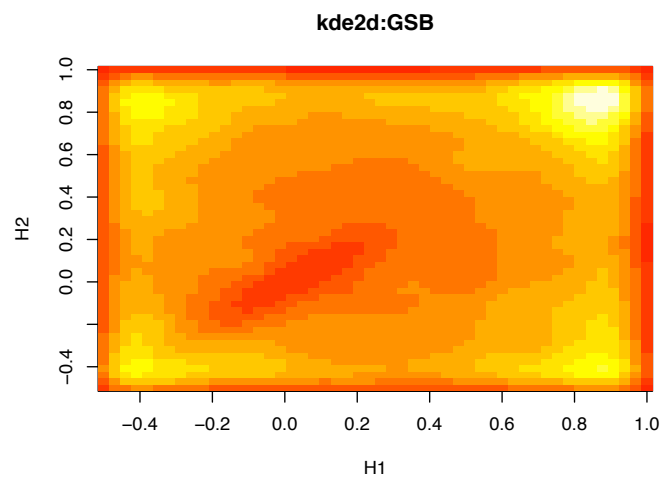


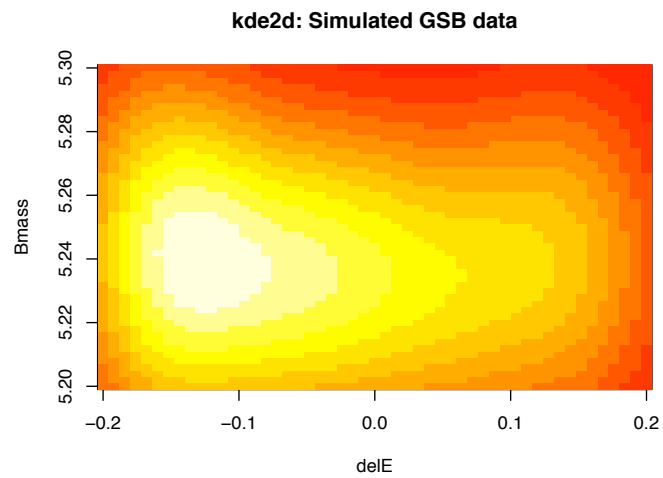
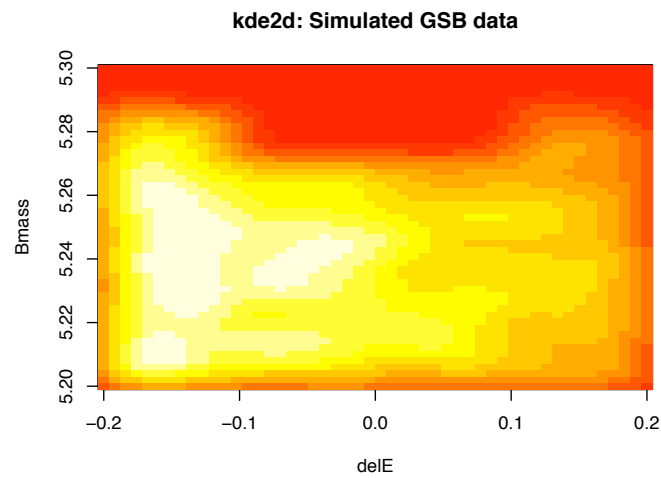
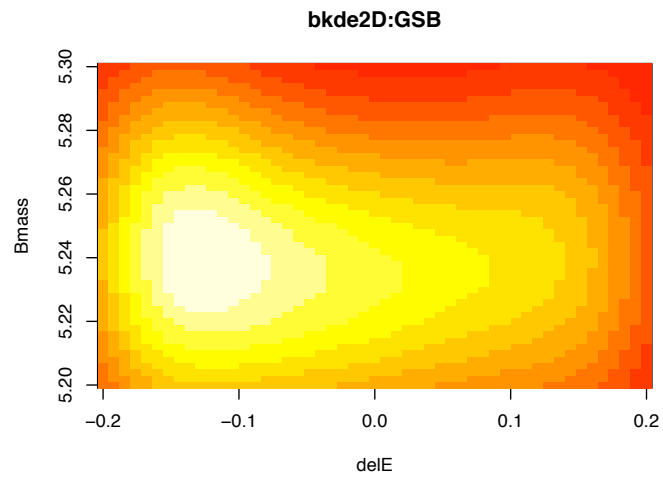
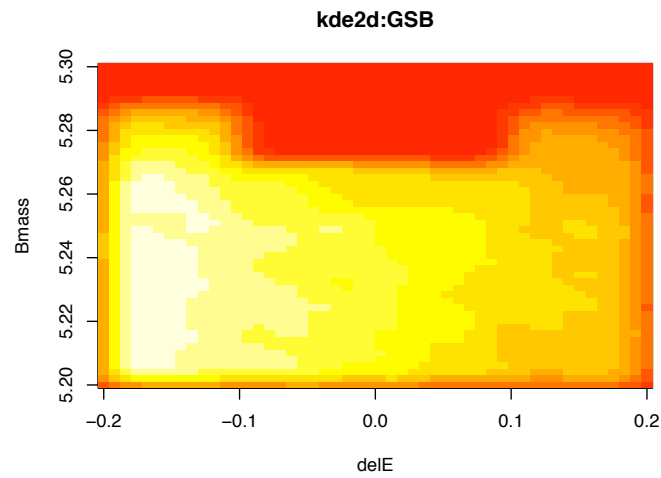


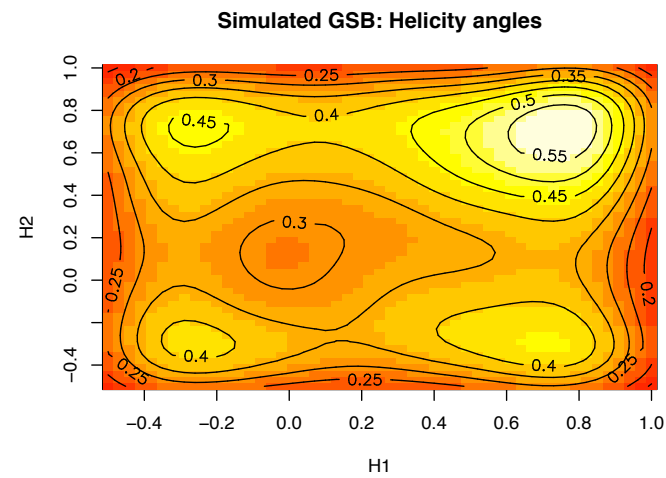
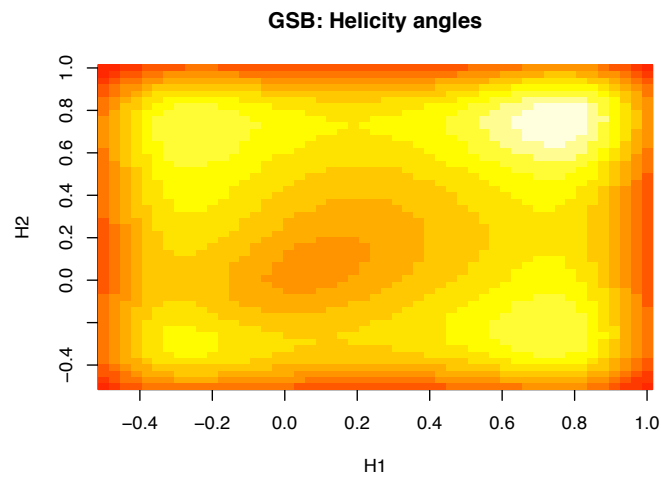
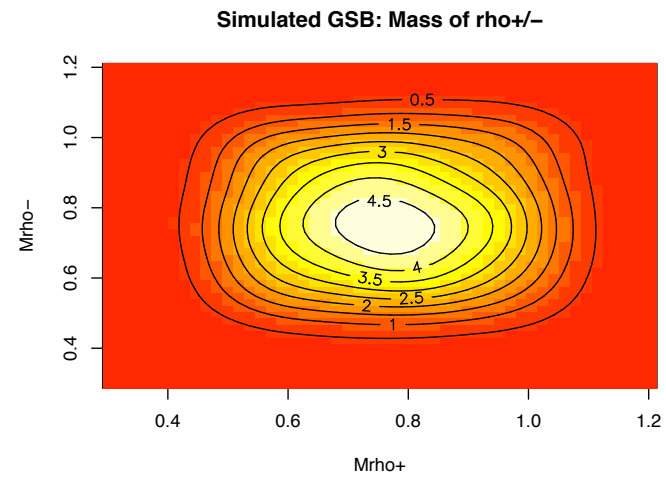
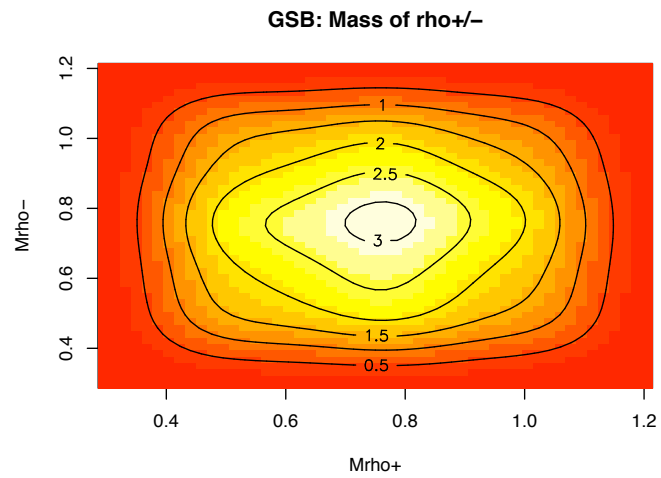






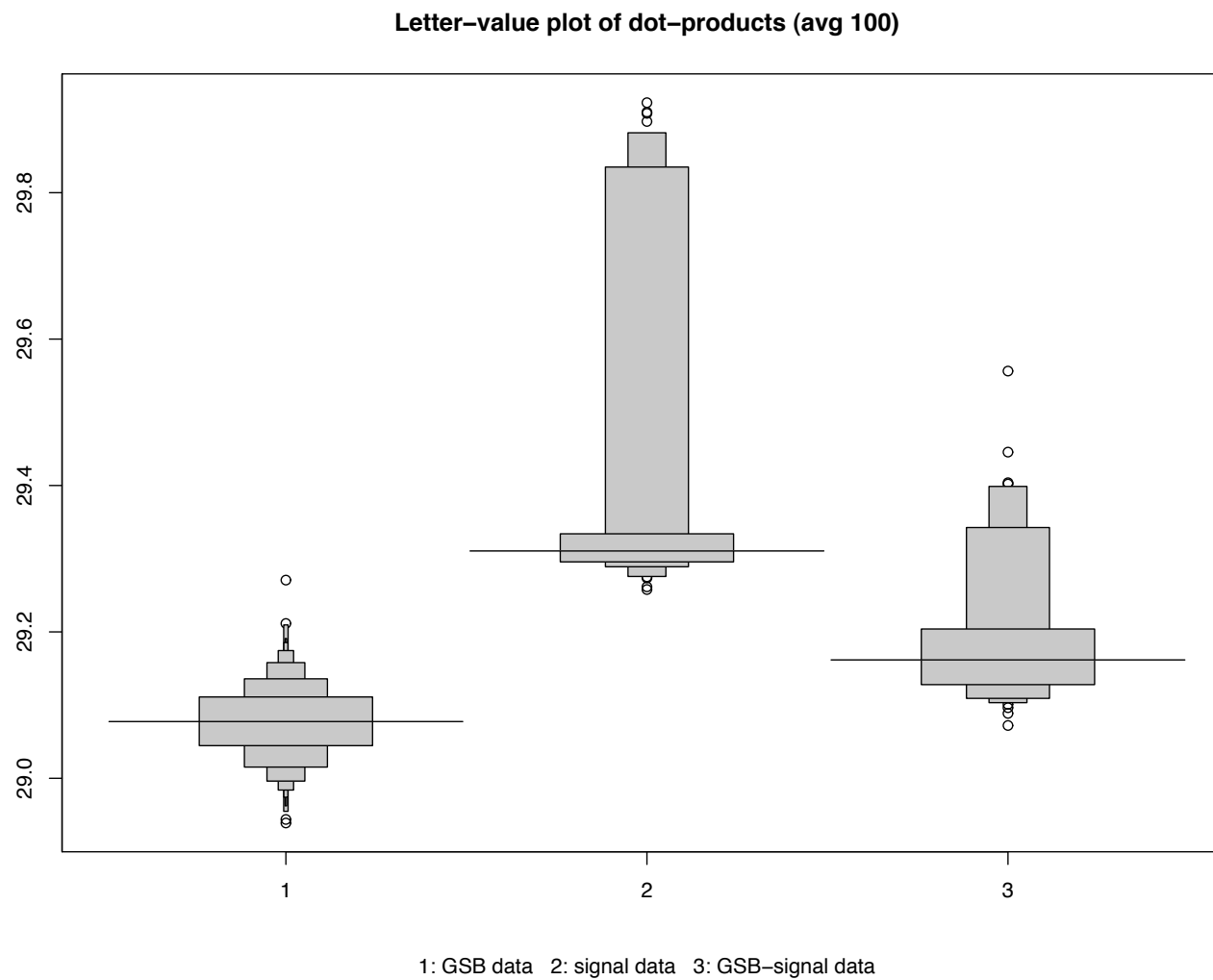






Initial attempts to identify target events:

- Average variables on 100 events at a time
- Dot-product each 100-average vector with subsequent vector
- *Does distribution of  $v'_1 v_2$  for GSB data look different from distribution of  $u'_1 u_2$  for signal data?*
- *Does distribution of  $v'_1 u_2$  for GSB  $v_1$  and signal  $v_2$  look different from distribution of  $v'_1 v_2$  for GSB data?*



Are 7 features sufficient to identify  $\rho - \rho$  events?

Suppose  $K$  *independent* features of particle A:

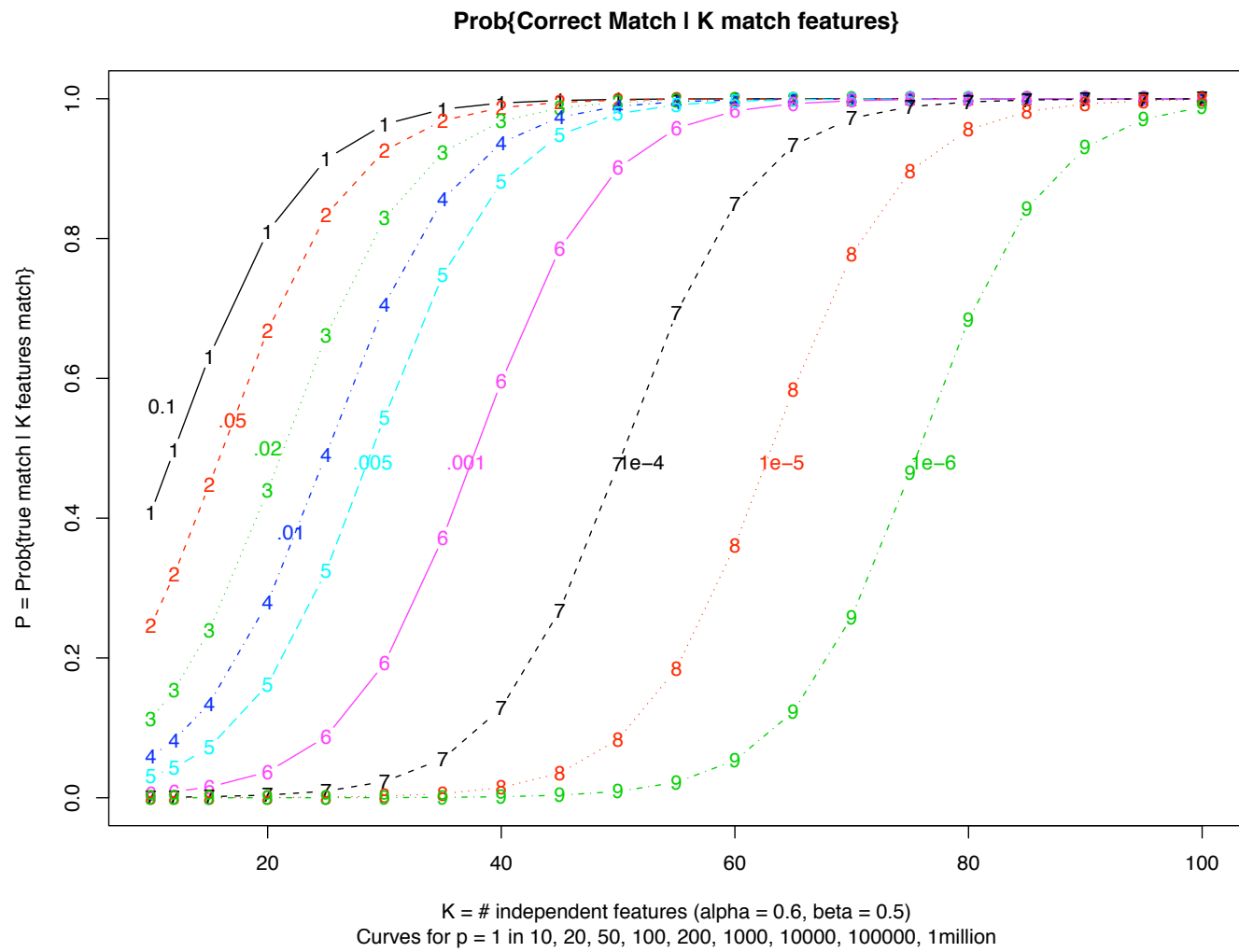
- $\alpha_i = P\{ \text{feature } i \text{ indicates A} \mid \text{Particle A} \}$
- $\beta_i = P\{ \text{feature } i \text{ does not indicate A} \mid \text{Not Particle A} \}$
- $p = P\{ \text{Particle A} \}$  (frequency of occurrence)

Bayes rule:

$$\begin{aligned} P(A) &= P\{\text{Particle A} \mid \text{All features indicate A}\} \\ &= \frac{p \prod_{i=1}^K \alpha_i}{p \prod_{i=1}^K \alpha_i + (1-p) \prod_{i=1}^K (1-\beta_i)} \end{aligned}$$

How large must  $K$  be, for  $P(A)$  to be ‘large’?

Consider  $\alpha = 0.6$ ,  $\beta = 0.5$  (tiny  $P(A)$  when  $\beta < 0.5$ )





### Derived variables: Simulated data

- Graphical comparisons of GSB (real grand side band) data and simulated background data appear reasonably close
- Simulated signal (rho true) data set has 5,913 events; all but 20 have `Bmass` > 5.265
- Simulated noise (background) data set has 9,937 'B' events (either  $B^+B^-$  or  $B^0\bar{B}^0$ ), but only 2,404 have `Bmass` > 5.265
- Compare a random selection of 2,404 signal events with 2,404 background events
- For 5,913 signal events:
  - `delE` has mean 0.012, SD 0.0580
  - `Bmass` has mean 5.279, SD 0.0034
  - Bivariate plot of normalized `Bmass` versus normalized `delE` is roughly circular

- Mrho+, Mrho- have mean 0.784, SD 0.1100
- Bivariate plot of normalized Mrho- versus normalized Mrho+ is somewhat circular

Bivariate density plots suggest two useful variables:

- EB-rad = [Euclidean distance of (delE, Bmass) from (0, 5.28)] <sup>$\frac{1}{3}$</sup>   
(after standardizing delE and Bmass by their SDs)
- RR-rad = [distance of (Mrho+, Mrho-) from (0.784, 0.784)] <sup>$\frac{1}{3}$</sup>   
(after standardizing Mrho+ and Mrho- by SD = 0.1100)

Can 4 features indicate regions of high signal density?

Define three intervals of roughly equal density of signal events for each variable:

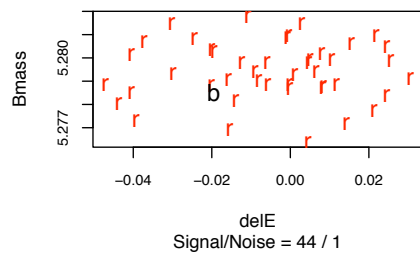
- `delE/Bmass` transformed radius: (0, 0.88, 1.28, 3.1)
- `Mrho+/Mrho-` transformed radius: (0, 0.88, 1.28, 2.8)
- H1: (-0.5, -0.016, 0.457, 1.0)
- H2: (-0.5, -0.004, 0.457, 1.0)

If variables are independent and coverage is uniform, expect  $2404/81 \approx 30$  events per cell

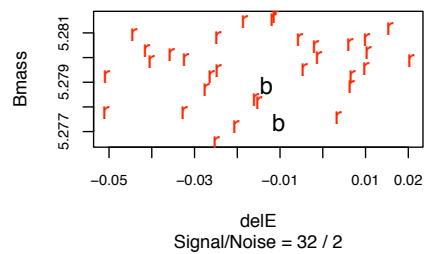
Plot `Bmass` versus `delE` for all 81 combinations

H1 = 1 H2 = 1

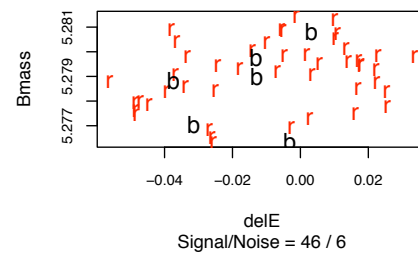
**BErad = 1 Mrad = 1**



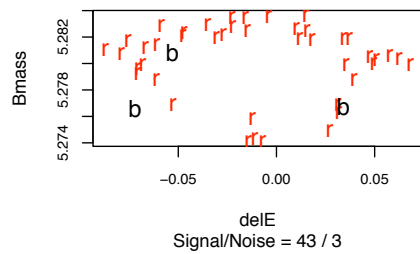
**BErad = 1 Mrad = 2**



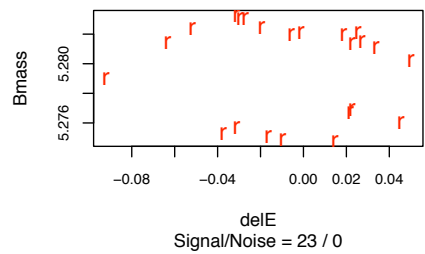
**BErad = 1 Mrad = 3**



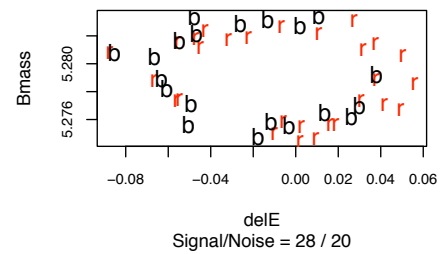
**BErad = 2 Mrad = 1**



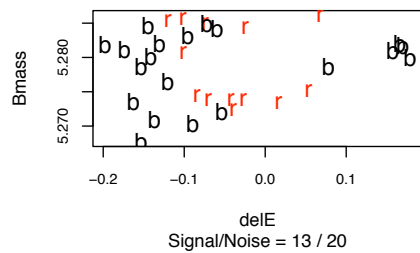
**BErad = 2 Mrad = 2**



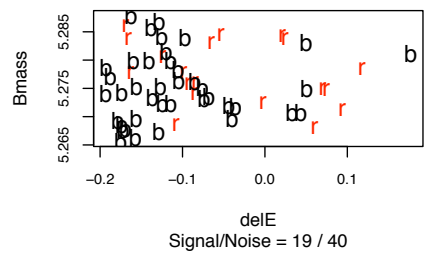
**BErad = 2 Mrad = 3**



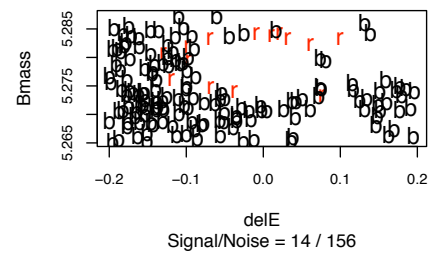
**BErad = 3 Mrad = 1**



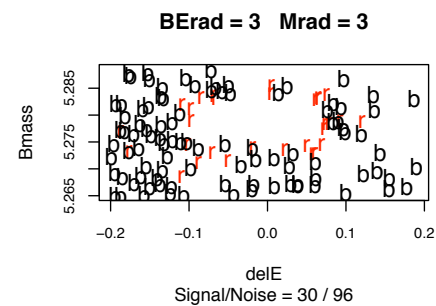
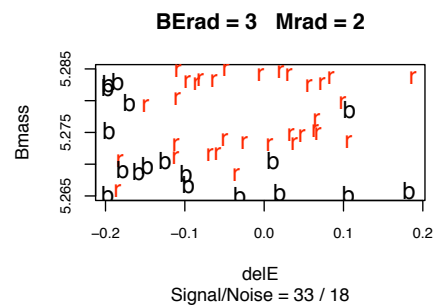
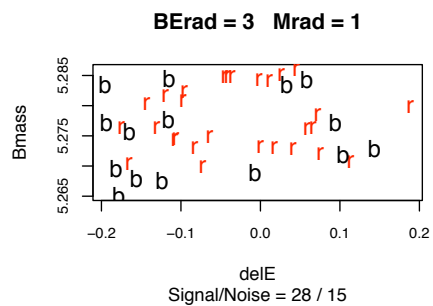
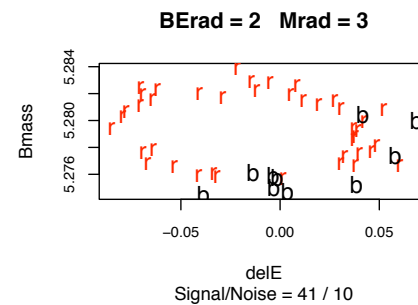
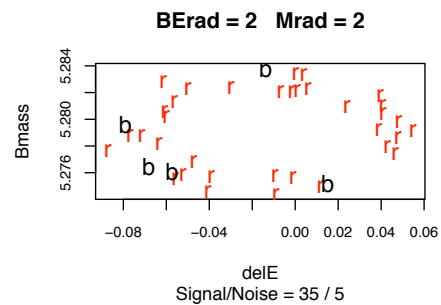
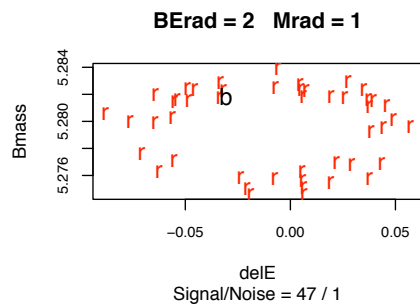
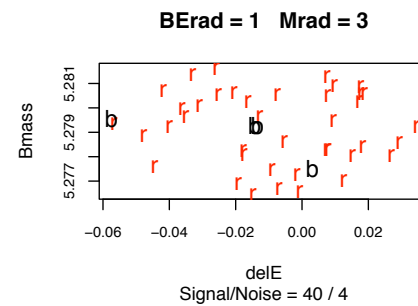
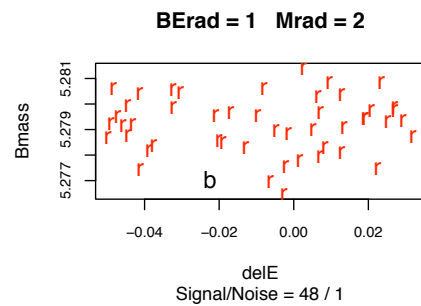
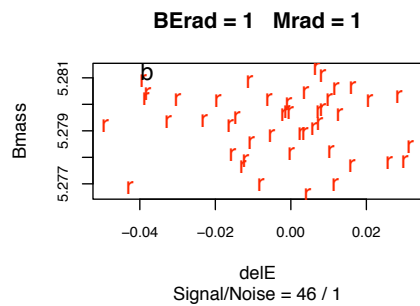
**BErad = 3 Mrad = 2**



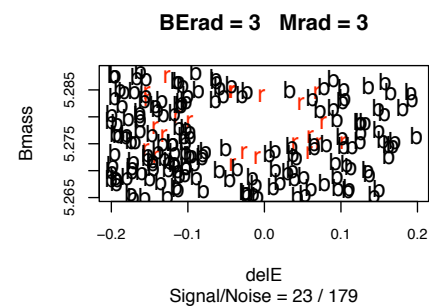
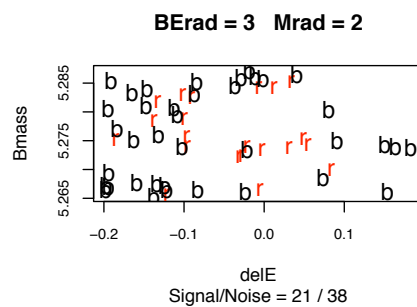
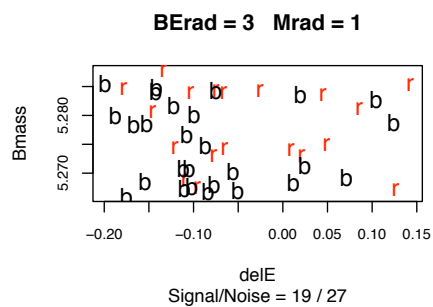
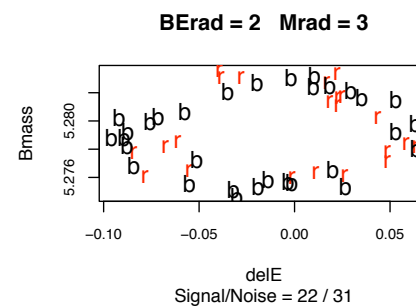
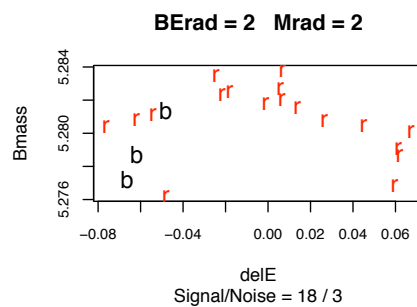
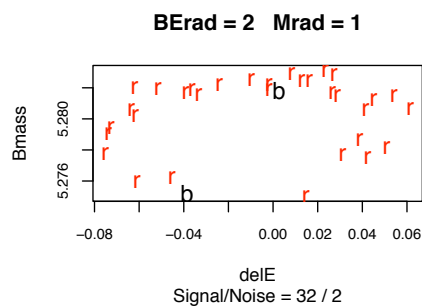
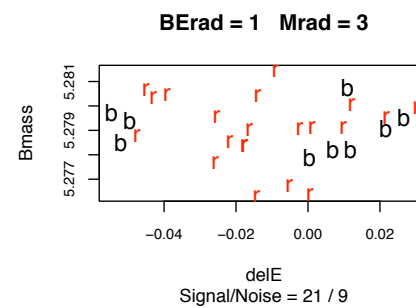
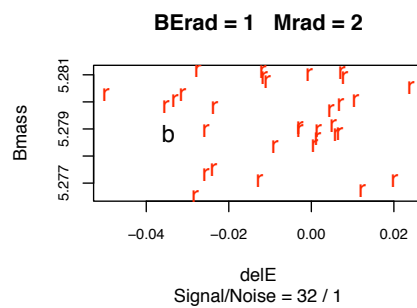
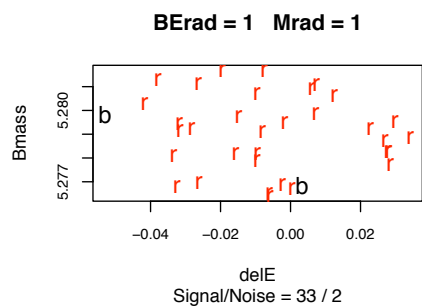
**BErad = 3 Mrad = 3**



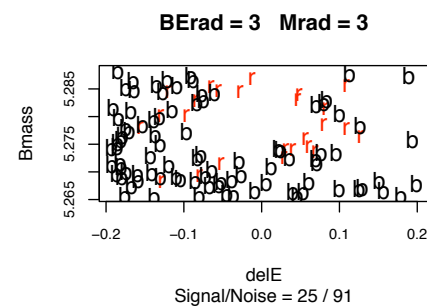
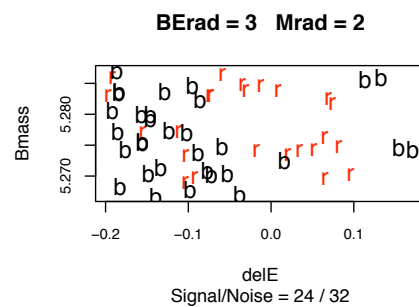
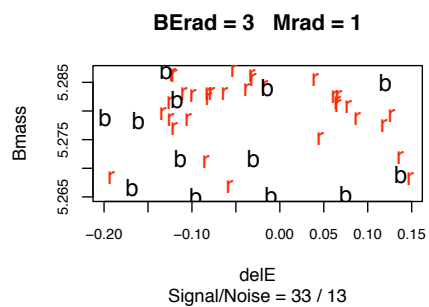
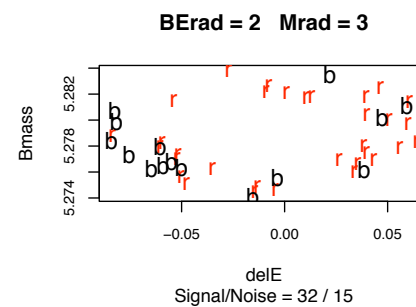
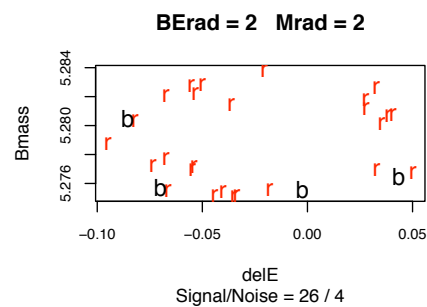
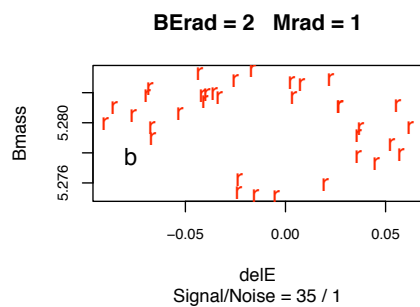
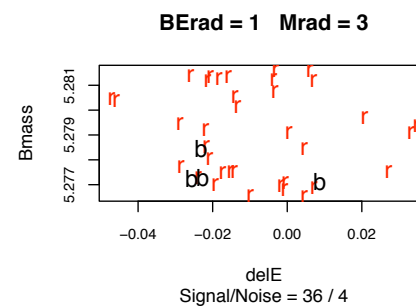
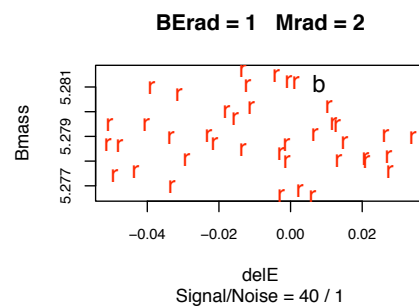
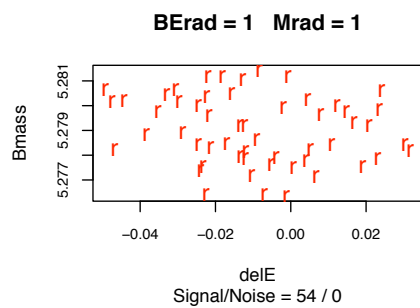
H1 = 1 H2 = 2



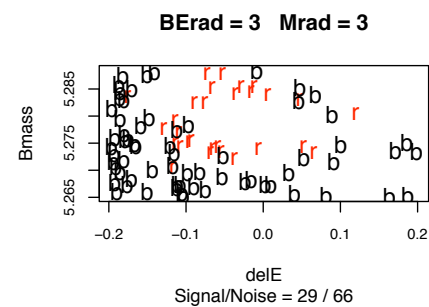
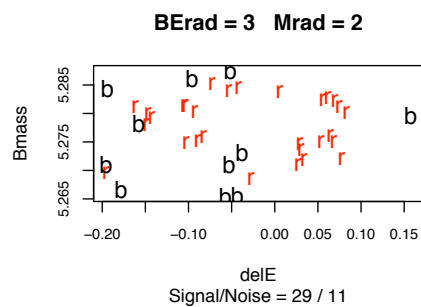
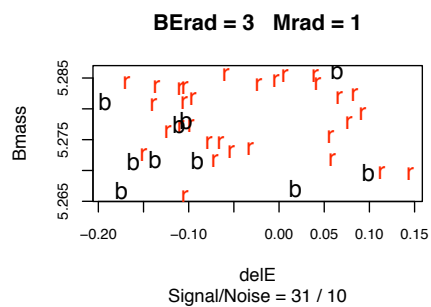
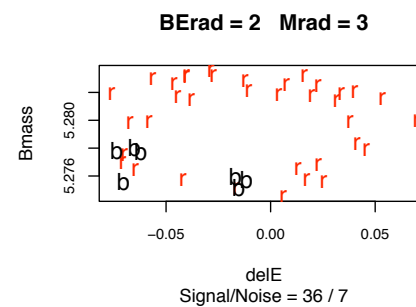
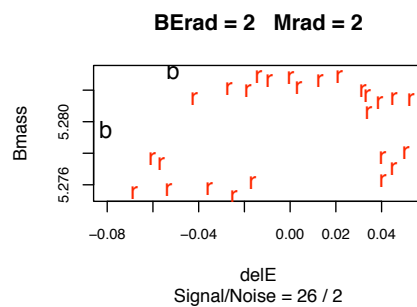
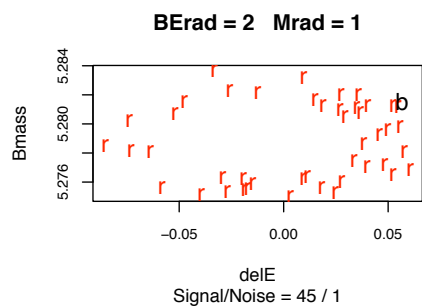
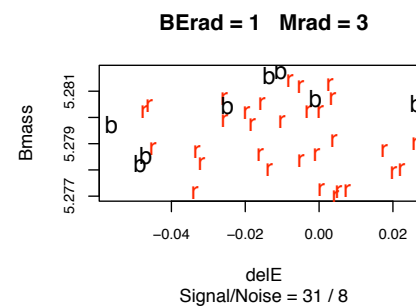
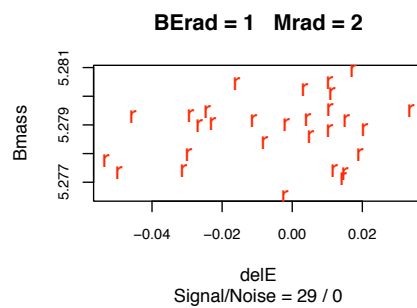
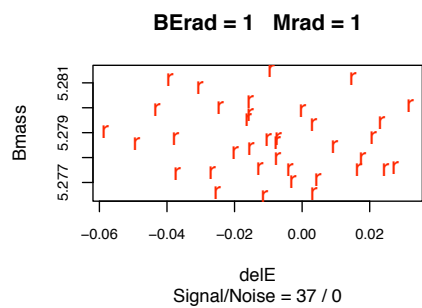
H1 = 1 H2 = 3



$$H1 = 2 \quad H2 = 1$$

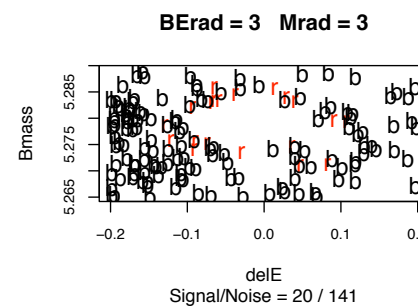
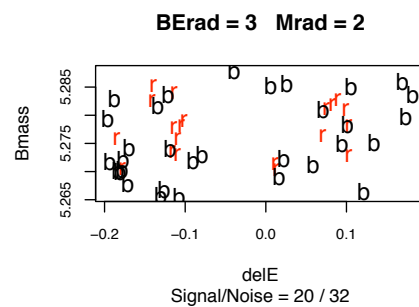
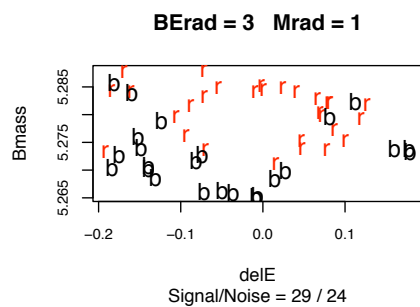
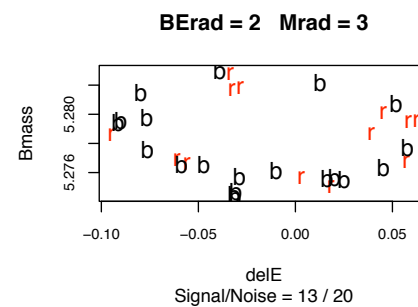
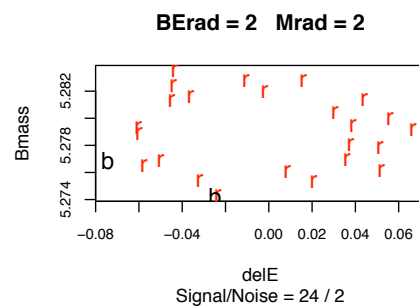
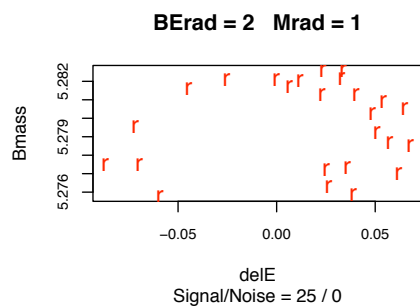
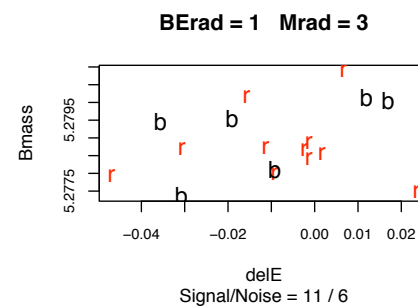
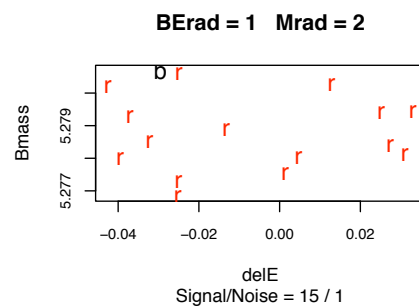
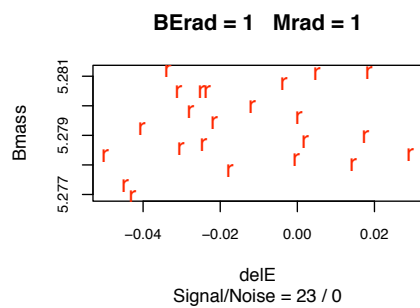


$$H1 = 2 \quad H2 = 2$$

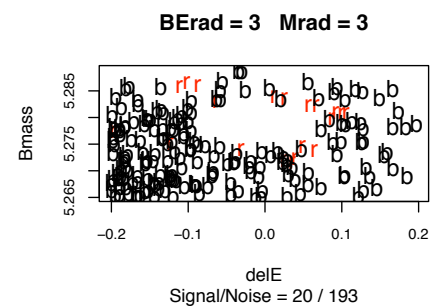
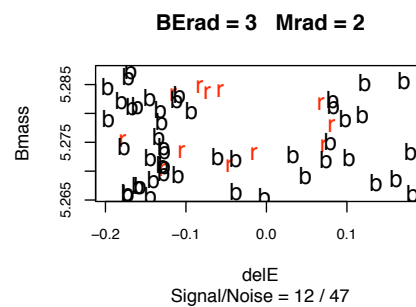
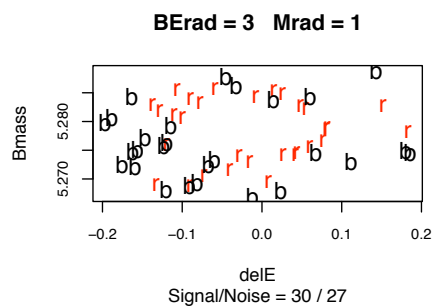
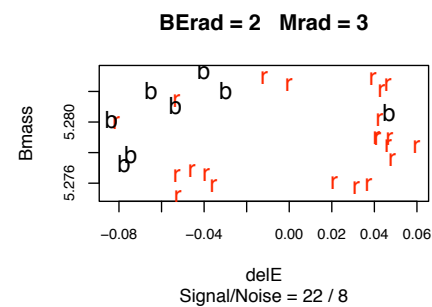
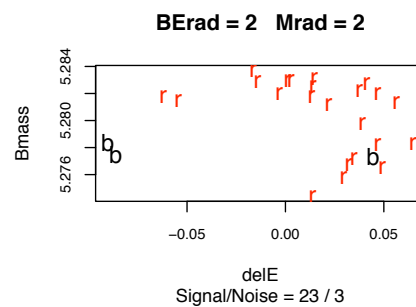
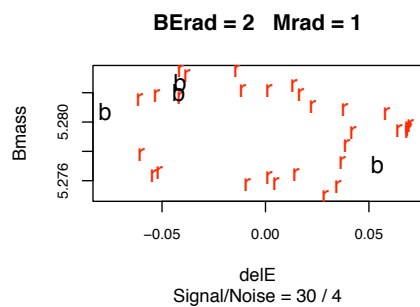
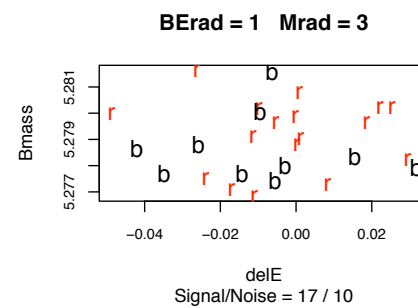
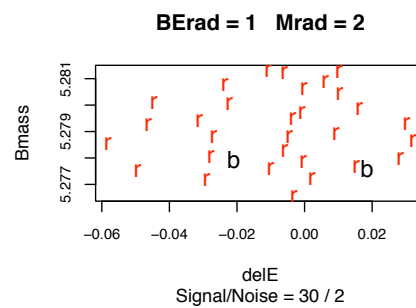
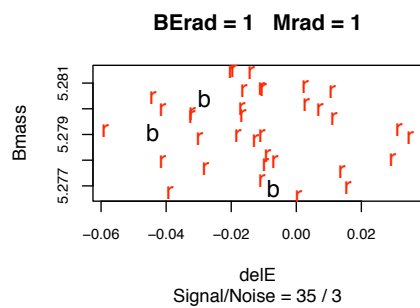




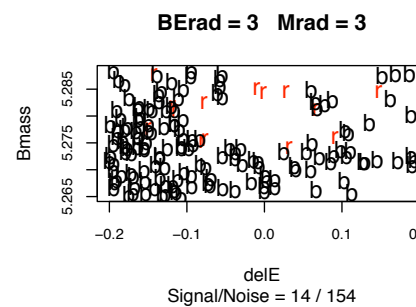
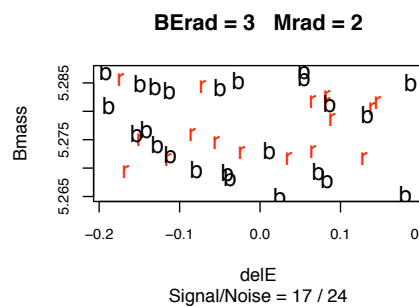
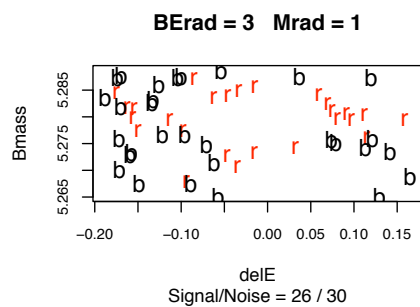
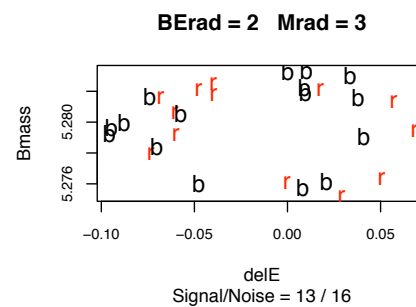
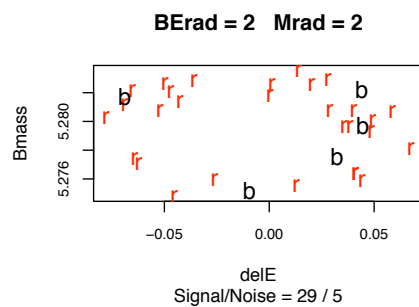
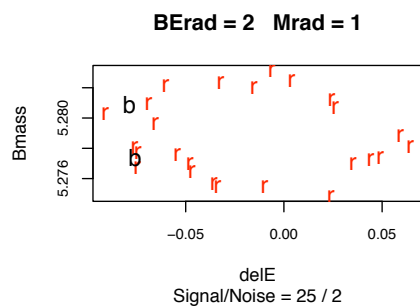
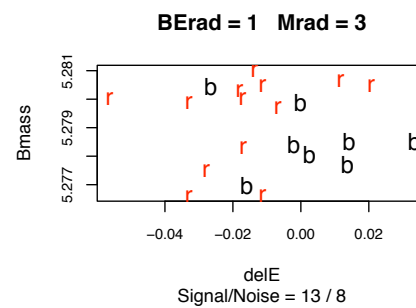
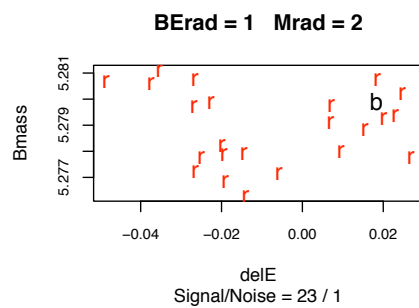
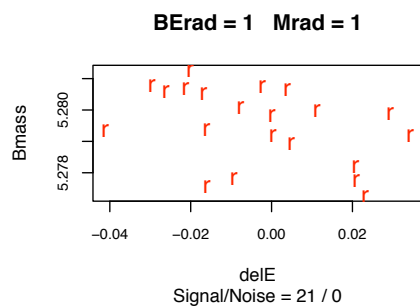
$$H1 = 2 \quad H2 = 3$$



H1 = 3 H2 = 1

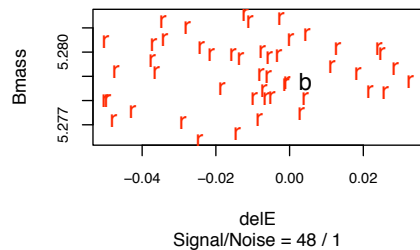


H1 = 3 H2 = 2

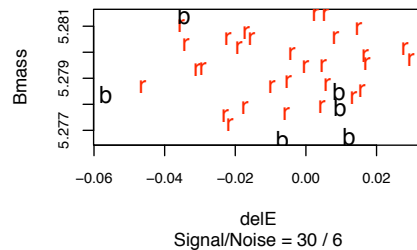


H1 = 3 H2 = 3

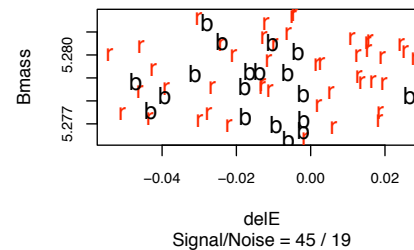
**BErad = 1 Mrad = 1**



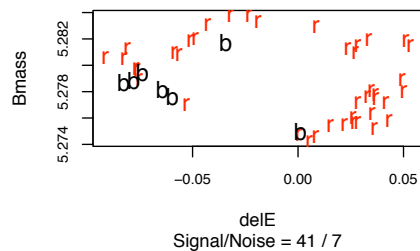
**BErad = 1 Mrad = 2**



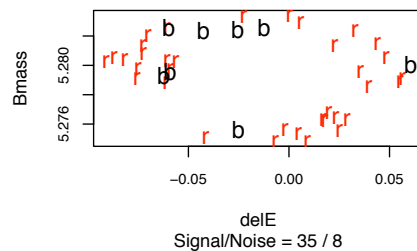
**BErad = 1 Mrad = 3**



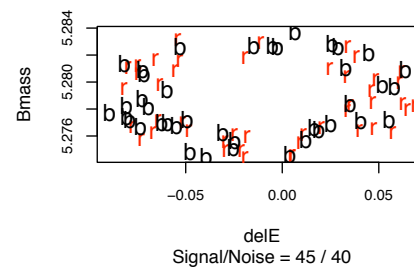
**BErad = 2 Mrad = 1**



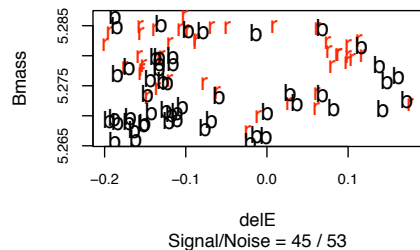
**BErad = 2 Mrad = 2**



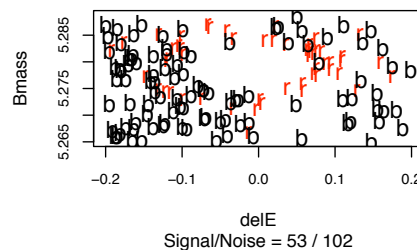
**BErad = 2 Mrad = 3**



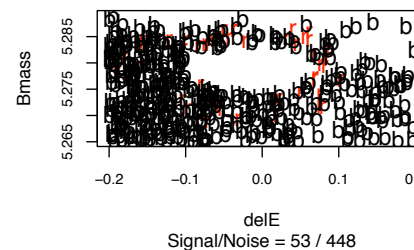
**BErad = 3 Mrad = 1**



**BErad = 3 Mrad = 2**



**BErad = 3 Mrad = 3**



Which of the 81 “signal-to-noise” ratios are “significant”?

Poisson test of rates:  $r\text{-count} \sim \text{Bi}(r\text{-count} + b\text{-count}, 0.5)$

‘Significant’ p-values by FDR at 0.05 (0.01) for 50 (47) ratios whose p-values are below 0.013 (0.008) in these regions:

- whenever point is close to center of (**delE,Bmass**) plot AND is close to center of (**Mrho+,Mrho-**) plot (i.e.,  $(radius)^{\frac{1}{3}} < 1.28$ )
- whenever point is very close to center of one plot but distant in second plot, if H1 and H2 are close to their means

Tabled entry: r-count/b-count, for each (EB-rad, RR-rad) region

	H2 = 1			H2 = 2			H2 = 3		
H1=1	44/1	32/2	46/6	46/1	48/1	40/4	33/2	32/1	21/9
	43/3	23/0	—	47/1	35/5	41/10	32/2	18/3	—
	—	—	—	—	33/18	—	—	—	—
H1=2	54/0	40/1	36/4	37/0	29/0	31/8	23/0	15/1	—
	35/1	26/4	32/15	45/1	26/2	36/7	25/0	24/2	—
	33/13	—	—	31/10	29/11	—	—	—	—
H1=3	35/3	30/2	—	21/0	23/1	—	48/1	30/6	45/19
	30/4	23/3	22/8	25/2	29/5	—	41/7	35/8	—
	—	—	—	—	—	—	—	—	—

## Summary and future work

- New data types/structures lead to advances in science
- Information age  $\Rightarrow$  Excellent opportunities for collaborations among statisticians, computer scientists, engineers
- Streaming data require:
  - much pre-processing to be interpretable
  - much summarization so they can be displayed
  - fast, **scalable** processing algorithms
- Streaming data offer new challenges to statisticians:
  - data acquisition, storage, distribution
  - fast algorithms and meaningful displays
  - better combinations of classical, robust analyses
- We still need exploratory plots:  
detecting “exotic” requires characterizing “typical”

- EDA helps to identify natural "units" for study
- We need new tools & displays for streaming data, but ...
- Displays will be monitored by non-statisticians, so interpretation must be clear:

*"Churchill Eisenhart ... defined practical power as the product of the mathematical power by the probability that the procedure will be used. A compact procedure may well be used so much more often as to more than compensate for its loss of mathematical power."*

— J.W. Tukey, "A Quick, Compact, Two-Sample Test to Duckworth's Specifications," *Technometrics* 1(1), p.32



## References

Benjamini, Y.; Hochberg, Y. (1995): “Controlling the false discovery rate...”, *JRSSB* 57: 289–300.

Marchette, David J. (2001), *Computer Intrusion Detection and Network Monitoring*, Springer, New York.

KK + EJW (2006), “Visualizing ‘typical’ and ‘exotic’ Internet traffic data,” *CSDA* 50: 3721-43.

Tukey, J.W. (1962), “The future of data analysis,” *AMS* 1962; 1–67.

*Scientific American*, “The Future of Physics,” Feb 2008.